New methods for seasonal adjustment

Dissertation
zur Erlangung des Doktorgrades
Dr. rer. pol.
der Fakultät für Wirtschaftswissenschaften
der Universität Duisburg-Essen

vorgelegt von
Daniel Ollech

Schleswig, Deutschland

Betreuer:
Prof. Dr. Christoph Hanck
Fakultät für Wirtschaftswissenschaften
Lehrstuhl für Ökonometrie

November 2024

Gutachter

Prof. Dr. Christoph Hanck
Prof. Dr. Vasyl Golosnoy

Tag der mündlichen Prüfung

19. August 2025

For my family. Especially Oma.

DOI: 10.17185/duepublico/84340
URN: urn:nbn:de:hbz:465-20251001-080150-7

Acknowledgements

I would like to thank my supervisor, Christoph Hanck, for readily helping with anything I needed, providing constructive feedback and giving me freedom in developing this research.

I am also indebted to my co-authors, Karsten Webel and Luca Nocciola, for their collaboration and invaluable insights. Working alongside them has been an enriching experience, and I have learned a lot from our joint efforts.

A special acknowledgment is extended to my colleagues at the Deutsche Bundesbank and the broader seasonal adjustment community worldwide. All of our discussions have significantly contributed to the advancements presented in this work.

Above all, I want to thank my family. To my grandmother Helga, whose persistent inquiries about the progress of this thesis have been a source of motivation. To my grandfather, Willy, for instilling in me the understanding that it is important to make an effort. To my parents, Adriane and Thomas, for giving me the liberty to find my own path in life. To my wife, Jaqueline, for keeping on making me happy. And to my children, Carlotta, Cleo and Levia, for being here.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Seasonal adjustment aims to remove predictable seasonal fluctuations and other calendar-related variations from a time series. In the context of economics Hylleberg (1992, p. 4) defines seasonality as "[...] the systematic, although not necessarily regular, intra-year movement caused by the changes of the weather, the calendar, and timing of decisions, directly or indirectly through the production and consumption decisions made by the agents of the economy".

The removal of seasonal and calendar effects allows analysts, economists, and statisticians to more clearly identify the business cycle and other non-seasonal effects, such as strikes or crises. By adjusting for seasonality, it becomes easier to compare data across different time periods and to make more timely economic assessments and policy decisions. The usual alternative, year-on-year growth-rates, show underlying developments with a time lag of about half a year (cf. Dagum and Mazzi, 2018). Moreover, these growth rates may be skewed by differences in the calendar constellation, for instance a discrepancy in the number of working days between the periods under comparison. To address these issues, seasonal adjustment techniques are employed to provide a more immediate and undistorted view of economic developments.

The collection of papers presented here adds to the literature on seasonal adjustment and time series econometrics by introducing contributions in two primary areas. Historically, the development and application of seasonal adjustment methods have predominantly focused on time series data with lower frequencies, such as monthly and quarterly observations. This work, however, develops and analyses methods to seasonally adjust higher frequency time series. Most prominently, a procedure is developed to adjust daily time series.

Furthermore, the second area of contribution lies in the analysis and enhancement of diagnostic tools that are instrumental in the seasonal adjustment process for both lower and higher frequency time series. This includes the examination and refinement of seasonality tests, which are pivotal in detecting seasonal patterns within the data. Additionally, the research delves into the realm of unit root tests, which are relevant in many areas of econometrics, and essential for the RegARIMA model based estimation of calendar effects. Collectively, these methodological advancements and diagnostic improvements not only facilitate more accurate seasonal adjustments but also contribute to the robustness and precision of time series econometrics as a whole.

Let us assume that a time series Y_t can be decomposed into its unobservable components following a basic time series model:

$$Y_t = T_t + S_t + C_t + I_t, \tag{1.1}$$

where T_t represents the trend-cycle component, encompassing the medium- and long-term progression of the series. The seasonal component S_t reflects the pattern recurring annually, and the calendar component C_t accounts for non-periodic variations associated with the calendar, such as the number of working days in a month or the timing of moving holidays such as Easter. The irregular component I_t captures all other short-term dependencies and stochastic fluctuations.

For monthly and quarterly time series, the ESS guidelines on seasonal adjustment - in the previous 2015 edition - provided by Eurostat (Eurostat, 2015) recommends specific methods for use. However, these methods, implemented as X-13ARIMA-SEATS and TRAMO/SEATS, are not designed to handle time series with higher frequencies. This raises the question: How can higher frequency time series be seasonally adjusted? Until recently, there were no established methods specifically designed to address the complexities of daily time series.

The **Second Chapter**[1] describes the Daily Seasonal Adjustment (DSA) procedure, which is tailored to the unique characteristics of daily time series. The basic time series model, as described by Equation 1.1, is extended to accommodate multiple periodic effects. Daily time series exhibit annual seasonality ($S_t^{(365)}$), akin to that in monthly and quarterly series, as well as weekly ($S_t^{(7)}$) and, for certain economic variables, monthly ($S_t^{(31)}$) recurring effects. The extended time series model is thus given by:

$$Y_t = T_t + S_t^{(7)} + S_t^{(31)} + S_t^{(365)} + C_t + I_t, \quad \forall t = 1, \dots, N. \tag{1.2}$$

The DSA method sequentially adjusts for seasonal and calendar effects by combining a RegARIMA model with the STL (Seasonal and Trend decomposition using Loess,

[1]Published as Ollech, D. (2021). Seasonal adjustment of daily time series. Journal of Time Series Econometrics, 13(2):235-264. https://doi.org/10.1515/jtse-2020-0028

Cleveland et al., 1990) seasonal adjustment method. In order to facilitate this sequential estimation approach, DSA handles the necessary data transformations within each step, e.g. it extends the time series to encompass 31 days per month for the adjustment of the day-of-the-month.

The DSA procedure is evaluated using a set of daily time series as well as simulated time series. The simulation algorithm, that is described in the paper, has been implemented in the R package {`tssim`}, which has been downloaded more than 10,000 times since June, 2021.

The DSA procedure is implemented in the R package {`dsa`}, which has been downloaded over 38,000 times since its release in 2018. Additionally, it has been incorporated into the commercial software EViews starting from version 13. The current ESS guidelines 2024 edition (Eurostat, 2024) include DSA as a recommended option.

The **Third Chapter**[2] complements the previous chapter in that it discusses the seasonal adjustment of hourly, daily and weekly time series - the procedures to seasonally adjust weekly and hourly time series presented in the paper are novel contributions. The COVID-19 pandemic heightened the demand for detailed and promptly available data, as economists and analysts grappled with assessing the economy's condition, particularly during and following the lockdown periods. As these time series are meant to be analysed at intra-annual intervals, a quality seasonal and calendar adjustment is paramount. Accordingly, this paper conducts a detailed analysis and adjustment of a set of high-frequency time series that continue to contribute to the German weekly activity index. This index, published by the Deutsche Bundesbank (Eraslan and Götz, 2020), monitors real economic activity in Germany and has a strong correlation with the Gross Domestic Product (GDP).

Beyond analysing how to seasonally adjust higher frequency time series, the paper highlights the features of higher frequency time series and their relevance for seasonal and calendar adjustment. Such an inventory of characteristics is an important building block in evaluating seasonal adjustment methods and can guide the development of diagnostic tools. For example, seasonality tests are used to detect periodically recurring effects and thus, help to decide whether a time series should be seasonally adjusted, or if a seasonally adjusted time series still contains residual seasonality. The number of seasonality tests in the literature is vast, some econometric software contains more than one seasonality test. Yet, how can we know which seasonality tests to trust the most?

[2] Published as Ollech, D. (2023). Economic analysis using higher-frequency time series: challenges for seasonal adjustment. Empirical Economics, 64(3):1375-1398. https://doi.org/10.1007/s00181-022-02287-5

The **Fourth Chapter**[3] evaluates a set of seasonality tests included in JDemetra+, the seasonal adjustment software endorsed by Eurostat, using random forests. Specifically, the variable importance estimation based on conditional random forests - a variant of random forests that accounts for the high correlation of the seasonality test outcomes - is used to estimate the contribution of each individual seasonality tests to the correct classification into non-seasonal and seasonal time series. Consequently, it enables the quantification of the relative informational value of each test and establishes a hierarchy based on their informational content.

To obtain generalisable results, the research introduces an algorithm designed to generate time series from ARIMA models that are representative of a given set of time series, e.g. all series included in a time series database. This is used to simulate both seasonal and non-seasonal data, so that it is known, whether a given series is seasonal or not. Seasonality tests are computed for all simulated time series.

The set of seasonality tests is then evaluated in two ways. First, each test's accuracy with respect to identifying non-seasonal and seasonal time series is computed. Second, the variable importance estimation in conditional random forests are applied to the seasonality test results.

The variable importance metrics indicate that only two of the tests implemented in JDemetra+ are indispensable, as the others provide redundant information. The two tests are elegantly complementary: the $Friedman$ test is specifically designed to detect deterministic seasonality, whereas the QS test builds on seasonal autocorrelation, and thereby allows the seasonal pattern to gradually evolve over time.

The **Fifth Chapter**[4] builds upon the preceding analysis by developing an overall seasonality test. Drawing from an expanded collection of seasonality tests, a recursive feature elimination algorithm - leveraging the variable importance measure from conditional random forests - is employed to identify the most informative seasonality tests. At the end of each elimination round, the accuracy of the conditional random forests comprising of the (reduced) set of seasonality tests with respect to classifying into seasonal and non-seasonal series is monitored. The final set of seasonality tests contains all tests that, used together as input into the conditional random forests, yielded the highest accuracy.

Subsequently, a single conditional inference tree is constructed using the refined subset of tests to determine the optimal assembly of these tests. The overall test amalgamates a test for stable seasonality - utilising an extended version of the $Kruskal - Wallis$

[3]Published as Ollech, D. and Webel, K. (2023). A random forest-based approach to combining and ranking seasonality tests. Journal of Econometric Methods, 12(1):117-130. https://doi.org/10.1515/jem-2020-0020

[4]Available as Ollech, D. and Webel, K. (2018). An overall seasonality test based on recursive feature elimination in conditional random forests. In Proceedings of the 5th International Conference on Time Series and Forecasting:20-31. https://dx.doi.org/10.2139/ssrn.3721055

test - and tests that allow changing seasonality, namely two variants of the previously mentioned QS test.

The overall seasonality test, in addition to all individual seasonality tests discussed in this paper, have been implemented in the R package {`seastests`}, which has been downloaded close to 100,000 times since March, 2019.

The **Sixth Chapter**[5] applies the methodologies developed in chapters four and five to both non-seasonal and seasonal unit root tests.

Consider the autoregressive process of order one, AR(1):

$$Y_t = \phi Y_{t-1} + \epsilon_t, \tag{1.3}$$

where ϕ represents the dependence of the time series Y_t on its past values, and ϵ_t denotes a white noise error term.

A unit root is present when $\phi = 1$, rendering the time series non-stationary. Consequently, the system's response to shocks may be enduring, affecting the series level permanently. Such a series will not revert to a long-term mean, and its variance will progressively increase over time. The presence of unit roots is a critical consideration, as many conventional statistical techniques presuppose stationarity in the time series under analysis. Ignoring unit roots can lead to erroneous inferences and unreliable forecasts.

The landscape of unit root tests is diverse, with a plethora of tests designed for various data characteristics, such as time series with deterministic trends, heteroskedastic innovation processes, or structural breaks. In practice, selecting the optimal suite of tests to accommodate all relevant scenarios is a challenge. The chapter demonstrates that the precision of individual unit root tests in accurately differentiating between series with and without a unit root is generally modest. However, considerable improvements can be achieved by combining multiple unit root tests.

Employing a simulation and inference approach akin to that used in the preceding chapters, this study assesses an extensive array of non-seasonal and seasonal unit-root test variants. It identifies effective combinations that enhance the overall testing procedure. Because there is a dominantly performing seasonal unit root test, the misclassification accuracy for seasonal unit roots can only be reduced by less than a percentage point to a little more than 6%. For non-seasonal unit root tests, the combined approach markedly reduces the misclassification rate - from over 15% for the best-performing individual unit root tests to approximately 10% for a composite of tests - thereby offering a substantial improvement in the detection of non-seasonal unit roots.

The **Seventh Chapter** extends the scope of the QS test - a seasonality test that has been highlighted for its accuracy in chapters four and five - to accommodate time

[5]Published as Nocciola, L., Ollech, D., and Webel, K. (2021). Unit Root Test Combination via Random Forests. In *International Conference on Time Series and Forecasting* (pp. 31–46). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-14197-3_3

series data with non-integer seasonal frequencies, such as those encountered in weekly series. The inherent challenge with non-integer seasonal periods lies in the fact that the periodic effects do not align neatly with individual observations; instead, they influence adjacent data points. This dispersion of effects can lead to an attenuation of the seasonal autocorrelations upon which the QS test is predicated.

To address this issue, a novel adaptation of the QS test that incorporates a mechanism for including adjacent observations in the computations of the autocorrelation is proposed. This refinement ensures that the autocorrelations reflect the true periodicity of the data, even when the seasonality is of non-integer length. Furthermore, a correction factor for the test statistic is introduced, which compensates for the dampening effect on the autocorrelations. This adjustment allows the modified QS test to utilise the conventional null distribution, thereby preserving the test's original interpretability and statistical properties.

The dual contributions of this chapter are thus: firstly, the generalisation of the QS test enhances its applicability to a broader range of time series, and secondly, the derivation of the attenuation factor corrects the size distortions encountered for non-integer seasonal frequencies.

Bibliography

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6(1):3–73.

Dagum, E. B. and Mazzi, G. L. (2018). *Seasonal Adjustment: Objectives, Definitions, Costs and Benefits*, chapter 2, pages 33–48. European Commission.

Eraslan, S. and Götz, T. (2020). Weekly Activity Index for the German Economy. Available under URL `https://www.bundesbank.de/wai`.

Eurostat (2015). ESS Guidelines on Seasonal Adjustment. 2015 edition. Publications Office of the European Union, Luxembourg. ISBN 978-92-79-45176-8.

Eurostat (2024). ESS Guidelines on Seasonal Adjustment. 2024 edition. Publications Office of the European Union, Luxembourg. ISBN 978-92-68-15943-9.

Hylleberg, S. (1992). General Introduction. In Hylleberg, S., editor, *Modelling Seasonality*, pages 3–14. Oxford University Press.

CHAPTER 2

Seasonal Adjustment of Daily Time Series

Daniel Ollech

Abstract

Currently, the methods used by producers of official statistics do not facilitate the seasonal and calendar adjustment of daily time series, even though an increasing number of series with daily observations are available. The aim of this paper is the development of a procedure to estimate and adjust for periodically recurring systematic effects and the influence of moving holidays in time series with daily observations. To this end, an iterative STL based seasonal adjustment routine is combined with a RegARIMA model for the estimation of calendar and outlier effects. The procedure is illustrated and validated using a set of daily time series with different seasonal characteristics as well as simulated data. The developed procedure closes a gap by facilitating the seasonal and calendar adjustment of daily time series.

Published as: Ollech, D. (2021). Seasonal Adjustment of Daily Time Series. Journal of Time Series Econometrics, 13(2):235-264.

Keywords: Daily data; multiple Seasonality; STL; ARIMA

2.1 Introduction

Progress in information technology will increase the availability of time series with a daily or even higher periodicity that exhibit seasonal behaviour. Possible examples range from data on air pollution, web-search keywords, and road traffic data to economic variables such as online sales, retail prices, and the amount of banknotes in circulation. Often, new developments cannot be evaluated directly from these unadjusted data, as periodical influences may obscure underlying tendencies. Furthermore, periodic influences may be of interest themselves. An increasing number of banknotes towards the end of the year may be due to the annually recurring Christmas shopping related demand spikes instead of worrisome money hoarding. Seasonal adjustment[1] aims at detecting and eliminating these kinds of influences to enhance the interpretability of the data.

Currently, there exists no officially recommended method for seasonal and calendar adjustment of time series with daily observations (cf. Eurostat, 2015). The aim of this paper is to identify a suitable procedure to estimate and adjust for regularly and periodically recurring systematic effects as well as the influence of moving holidays for daily time series. The procedure should be flexible enough to estimate all seasonal patterns with different periodicities. At the same time, it should be sufficiently fast to allow the adjustment of a large number of series on a daily basis.

The well-known seasonal adjustment procedures TRAMO/SEATS and Census X-13, as implemented in Census X-13ARIMA-SEATS, TSW rev. 924 and JDemetra+ 2.2.3, do not allow for the adjustment of data with a higher than monthly frequency. Alternative methods have been developed that aim at estimating some or all of the seasonal effects in higher frequency time series. One of the first methods proposed was Akaike's BAYSEA procedure that is based upon Bayesian time series modelling and was developed as an alternative to X-11 (Akaike, 1980; Young, 1996). In theory, the approach is not restricted to monthly or quarterly observations. Yet, it can be demonstrated by simulation that BAYSEA does not yield plausible results for higher frequency data with a substantial intra-annual seasonal pattern, due to very slow convergence (Ollech, 2018).

Pierce et al. (1984) use a two-step model-based approach to seasonally adjust weekly monetary aggregates. This method has been developed further by Cleveland et al. (2018). Trigonometric functions are designed to capture any deterministic weekly seasonality and an ARIMA model is fitted to the residuals to obtain estimates for the remaining stochastic seasonality. Without further adaptations, simple seasonal ARIMA models do not capture all different seasonal frequencies that we observe for daily time series. Especially the simultaneous modelling of weekly and annually recurring patterns is challenging, as will be discussed later.

The software STAMP 8.2, which is designed to seasonally adjust data using structural time series modelling, assumes that daily time series have an intra-weekly pattern only (see Koopman et al., 2009). Harvey et al. (1997) employ structural models to adjust weekly monetary aggregates. For intra-monthly and intra-annual effects they suggest

[1] In the context of daily data, seasonal adjustment refers to the process of estimating and adjusting all periodically recurring systematic effects with a cycle length of less or equal one year. For daily data, these can typically be weekly, monthly and annually recurring patterns.

using either trigonometric or spline functions. Koopman and Ooms (2003, 2006) show that state space modelling can be used to analyse seasonal daily time series as well, though their main focus lies on short-term forecasting instead of seasonal adjustment as such.

Ladiray et al. (2018) describe an X-11-type algorithm for the seasonal adjustment of a series with two seasonalities, which thus can be used for some daily time series. The main idea is to iteratively estimate the seasonalities based on the trend-cycle adjusted series using period-wise seasonal moving average filters. The authors also discuss a SEATS-type algorithm that is based on a particular fractional ARIMA model, but currently cannot handle multiple seasonalities.

For the purposes of forecasting complex time series De Livera et al. (2011) introduce a state-space based model named TBATS (Trigonometric, Box-Cox transform, ARMA errors, Trend, and Seasonal components). It could be used for seasonal adjustment as well, though its main design-goal is the prediction of time series.

Dokumentov and Hyndman (2015) develop a regression based seasonal adjustment procedure, STR, that aims at providing automatic detection of multiple relevant seasonal cycles. However, the user has only limited influence on the outcome of this procedure. Depending on the complexity of the time series, the adjustment may also be computationally expensive. Section 2.5 includes an empirical comparison between STR, TBATS and the routine developed in this paper.

Semi-parametric STL provides another method, which is based on a locally weighted regression smoother (Loess). Compared to the methods discussed above, its main advantage is its high flexibility with respect to the frequency of the time series (Cleveland et al., 1990). Additionally, the fast computation of the STL algorithm makes it feasible to adjust different seasonal frequencies in an integrated iterative framework. Other authors have used STL for different applications involving seasonal time series. Verbesselt et al. (2010) develop a trend and seasonal change detection method that is based on STL and can be used for daily and other higher frequency time series. Thus, STL has properties that make it a suitable candidate for the seasonal adjustment of data with higher frequency.

We contribute to the existing literature by devising a seasonal adjustment routine for daily data. In addition to STL's limitations to single seasonal frequencies it cannot deal with calendar effects such as the influences of moving holidays. Therefore, we use STL as the basis of the seasonal adjustment and combine it with RegARIMA based estimations of calendar effects to obtain the daily seasonal adjustment (DSA) procedure.

The remainder of the paper is structured as follows: Section 2 presents key aspects of STL. Section 3 adapts STL to handle daily time series. Section 4 applies the procedure to a set of real world data and section 5 compares the results to those obtained using

STR and TBATS. Section 6 compares the procedures using simulated data. Section 7 concludes.

2.2 STL

Based on the well-known component model

$$Y_t = T_t + S_t + I_t \quad \forall \quad t = 1, ..., N \tag{2.1}$$

STL decomposes a time series (Y_t) into a trend-cycle (T_t), a seasonal (S_t) and an irregular component (I_t) using Loess regressions and moving averages (Cleveland et al., 1990). In Loess regressions, a weight is attached to each observation of the time series (Cleveland and Devlin, 1988). This weight is negatively related to the distance (in time) between a given observation and the value that is to be smoothed. If the distance is too large, the weight is zero. Thus, Loess regressions are local regressions because each value is regressed on a local neighbourhood of a linear or quadratic function of the (weighted) observations. For any point in time t, the weight $v_i(t)$ of the observation at t_i is given by

$$v_i(t) = \left[1 - \left(\frac{|t_i - t|}{\delta_\gamma(t)}\right)^3\right]^3 \tag{2.2}$$

with $\delta_\gamma(t) = |t_i - t_\gamma|$ being the distance between the γ^{th} farthest t_i and t. The smoothing parameter γ determines the number of (past and future) neighbouring observations included in the local regression. In STL, the user can specify the value of γ for the computation of the trend and seasonal component. The flexibility of the identified seasonal factor decreases with a higher value for γ. Correspondingly, the value of γ should be adapted to the variability of the observed seasonality (see also Section 2.4 for examples). As can be seen from Equation (2.2), the impact on the estimated seasonal component of increasing γ by one step depends inversely on the level of γ. Thus, changing γ from 9 to 13 has a similar impact to increasing it from 13 to 21.

Essentially, STL consists of an inner loop that is used for the decomposition of the series into the trend, seasonal and irregular component, and an outer loop for extreme value adjustment. The latter increases the robustness of the decomposition against outliers and thereby mitigates the influence of extraordinary events on the identification of the seasonal factors. STL can only adjust the seasonal pattern of a single frequency f at a time. The other seasonal patterns will be included in the irregular component.

2.2.1 Inner Loop

For the decomposition of Y_t, the inner loop goes through the following computations in the k-th iteration:

1. **Trend adjustment:**

$$Y_t - T_t^{\{k-1\}} = S_t^{\{k\}} + I_t^{\{k\}} \equiv TA_t^{\{k\}}, \text{ with } T_t^{\{0\}} = \mathbf{0}. \tag{2.3}$$

2. **Preliminary periodwise smoothing:** Each subseries of $TA_t^{\{k\}}$, i.e. each weekday, day of the month or day of the year for a daily time series or each month for a monthly time series, is smoothed by Loess to yield a preliminary seasonal factor $S_{pre,t}^{\{k\}}$. The γ for these regressions is the only parameter without a default in the STL procedure and thus has to be specified by the user (see Cleveland et al., 1990).

3. **Smoothing preliminary seasonal component:** A low-pass filter, composed of moving average- and Loess-regressions, is used to capture any low-frequency movements $L_t^{\{k\}}$ from $S_{pre,t}^{\{k\}}$.

4. **Obtaining seasonal component:**

$$S_t^{\{k\}} = S_{pre,t}^{\{k\}} - L_t^{\{k\}}. \tag{2.4}$$

5. **Seasonally adjusting the original time series:**

$$Y_t^{\{k\}} - S_t^{\{k\}} \equiv SA_t^{\{k\}}. \tag{2.5}$$

6. **Obtaining trend:** Loess is applied to $SA_t^{\{k\}}$ to yield $T_t^{\{k\}}$.

2.2.2 Outer Loop

Usually, only a few iterations of the inner loop are needed to ensure convergence of the results. By default, the number of iterations is set to 3 and the outer loop is omitted. Only if deemed necessary by the user, robustness weights for the Loess regressions of the inner loop are determined for each time point as a function of the size of the irregular component derived when the inner loop has reached its maximum number of iterations. The irregular component can be obtained by rearranging Equation (2.1):

$$I_t = Y_t - T_t - S_t. \tag{2.6}$$

Then, for the Loess regressions in the 2. and 6. part of the inner loop, the weights defined in Equation (2.2) are multiplied with the robustness weights given by Tukey's (1960) biweight function

$$\omega_t = \begin{cases} \left(1 - \left[\frac{|I_t|}{6 \cdot median(|\{I_t\}_{t=1}^N|)}\right]^2\right)^2 & \text{if } |I_t| < 6 \cdot median(|\{I_t\}_{t=1}^N|) \\ 0 & \text{else} \end{cases} \tag{2.7}$$

that strongly downweights highly irregular observations and is used in robust estimation.

2.3 Daily Seasonal Adjustment

Daily time series can contain different periodic patterns often resulting from regularities in the behaviour of people and economic conditions. This can be illustrated with the daily time series of the currency in circulation in Germany (see figure 2.1, a comprehensive discussion of the factors determining supply and demand of the currency in circulation can be found in Bartzsch et al., 2015 and Cabrero et al., 2009): if people primarily need money in cash to buy groceries for the weekend, the data will display a weekly pattern of increasing currency in circulation towards the end of the week. An increase around the end of the month of the circulating currency results from an increased withdrawal of cash following the payment of salaries. Finally, consumption increases due to Christmas, and thus the currency in circulation increases at the end of each year. Correspondingly, many daily time series are characterised by weekly, monthly and annually recurring patterns. In addition, travel activities around moving holidays might increase the demand for banknotes as well. Thus, the model described in Equation (2.1) has to be refined to incorporate intra-weekly, intra-monthly and intra-annual periodic patterns, $S_t^{(7)}$, $S_t^{(31)}$, and $S_t^{(365)}$, as well as moving holiday effects C_t:

$$Y_t = T_t + S_t^{(7)} + S_t^{(31)} + S_t^{(365)} + C_t + I_t. \tag{2.8}$$

For some series, the seasonal and calendar effects are not independent. The effect of fixed holidays for the electricity consumption in Germany depends on the weekday: as the weekday effect for the electricity consumption is influenced by how many people are working at a given day, the electricity consumption will be similar to a typical Sunday if, e.g., the labour day falls on a Sunday, while it drops considerably, if it falls on a working day. Another example is the air pollution in Europe. NO2 immissions are lower on Sundays, but this effect is more pronounced in winter than in summer. Thus, the weekday effect changes systematically over the year. These cross-seasonal effects can be included into Equation (2.8) as follows:

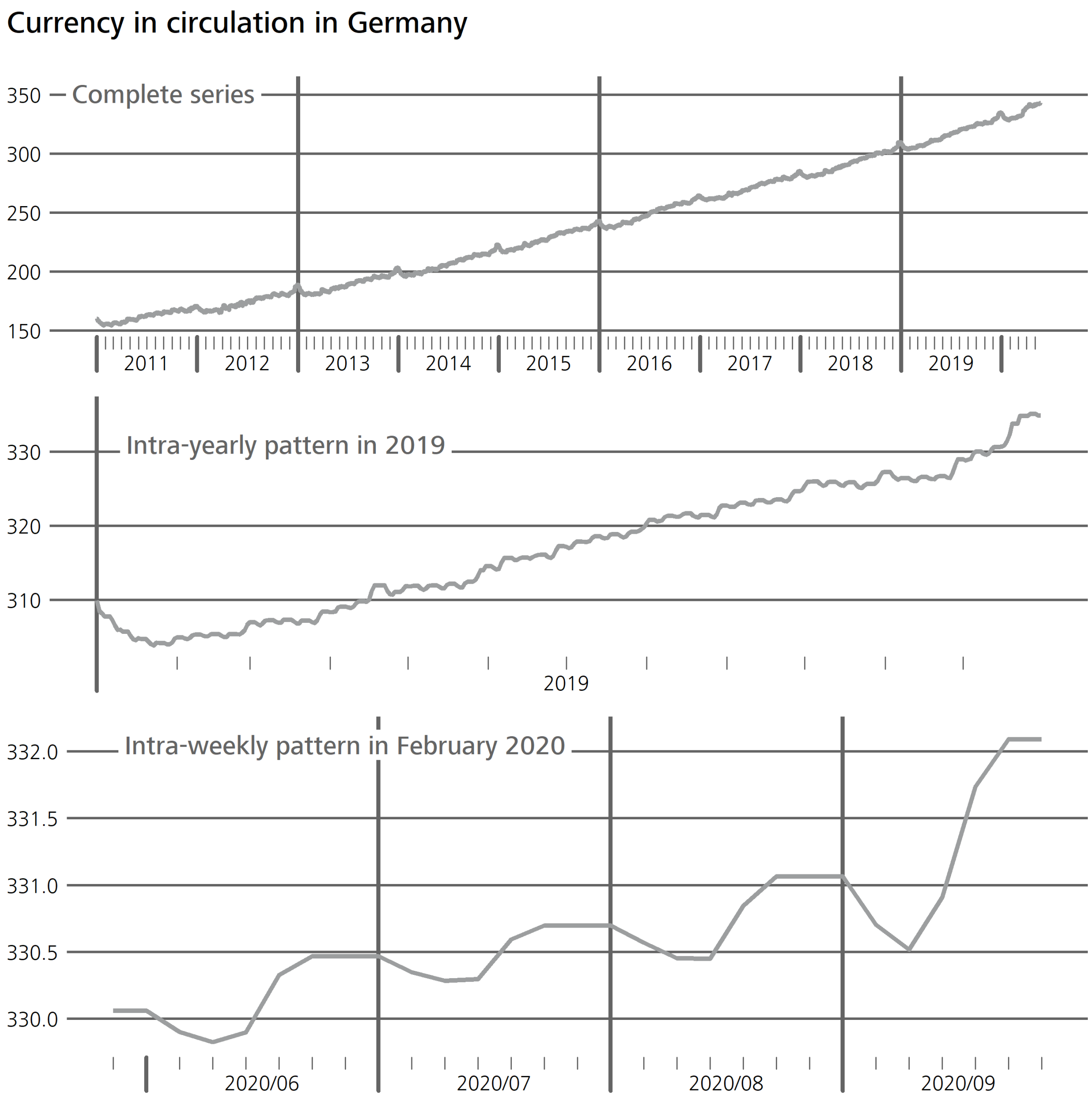

Figure 2.1: Currency in circulation in Germany, in billion Euro (see section 2.4.1 for details).

$$Y_t = T_t + S_t^{(7)} + S_t^{(31)} + S_t^{(365)} + C_t + S_t^{(7)} \times S_t^{(365)} + I_t. \tag{2.9}$$

STL adjusts each of the seasonal effects separately, and therefore has to be run three times in order to capture each $S_t^{(f)}\ \forall\ f \in (7, 31, 365)$.

The idea of iteratively adjusting several seasonal frequencies has already been put forward by Cleveland et al. (1990). In the case of daily time series, the main obstacle is the constraint of STL to have a constant period length, i.e. the number of observations has to be identical in each period. This is not unique to STL, though. For example, seasonal RegARIMA models usually face the same requirements. While this is unproblematic in the case of intra-weekly seasonality, the number of days per month and the number of days per year are not fixed. As a result, the period length has to be standardised, either by omitting a subset of the data or by artificial prolongation. Possible implementations are discussed below.

A further task is the development of a calendar and outlier adjustment routine together with the forecasting of seasonal factors. The latter is necessary for data producers providing seasonally adjusted figures applying a current adjustment revision policy, i.e. the policy of reidentifying time series components at set review periods and adjusting new data points by forecasted seasonal factors in the meantime (Eurostat, 2015). Outlier adjustment can be important for an unbiased estimation of the calendar effects as well as for the prediction of the future trajectory of the time series. The calendar and outlier adjustment needs to be integrated into the iterative seasonal adjustment.

As in X-13 and TRAMO/SEATS, the RegARIMA model could be the basis for all prior adjustments including estimating calendar and cross-seasonal effects. In the case of daily time series, several arguments can be made for changing this order and to begin with adjusting the intra-weekly movements: firstly, each year of data contains more than 50 observations of every weekday. The adjustment of the intra-weekly seasonality therefore tends to be robust against outliers and rare effects such as moving holidays even in the case of short series. Secondly, the intra-weekly, intra-monthly and intra-annual periodic movements necessitate different data modifications that are somewhat antagonistic: while the weekday effects need the full set of available data, the adjustment of the effect of the day of the year entails the standardisation of the length of the year to be either 365 or 366. If we started by estimating the effects of the moving holidays, there would be a risk that the estimated effects contain some of the weekday pattern.

Taking all of these considerations into account results in the following algorithm:

- Step I: Adjusting intra-weekly periodic effects with STL.
- Step II: Calendar-, cross-seasonal effects and outlier adjustment with RegArima.
- Step III: Adjusting intra-monthly periodic effects with STL.
- Step IV: Adjusting intra-annual periodic effects with STL.

It often will be the case that a given time series does not contain all of the seasonal and calendar components. This can be assessed by using seasonality tests, regression based statistics and spectral diagnostics. Then, it will be advisable to omit the respective steps.

2.3.1 Step I: Adjusting intra-weekly periodic effects with STL

Before estimating the intra-weekly seasonality with STL, the series may have to be pre-adjusted. This may include the interpolation of missing values or taking the logarithm of the data to stabilise the variance of the series.

For most of the computations in STL there exist sensible default values for the smoothing parameters (see Cleveland et al., 1990). Only for the smoothing of the period-wise sub-series $\gamma_{S^{(7)}}$, i.e. the number of observations included in the local regression to obtain the seasonal component, a specification is needed. In the case of intra-weekly seasonality, $\gamma_{S^{(7)}}$ has to be large enough so that the weekday effects do not get confounded by irregularities. Obviously, the parameter $\gamma_{S^{(7)}}$ should not be set too high either in order to capture changes in the seasonal pattern. In practice, often the period-specific raw seasonal factors, i.e. the sum of the estimated irregular and seasonal component, is inspected in order to determine an appropriate setting for $\gamma_{S^{(7)}}$.

As the time series has not been outlier adjusted yet, usually, the robust version of STL is preferable, so the number of iterations of the outer loop, which can be set by the user, is one or higher.

2.3.2 Step II: Calendar-, cross-seasonal effects and outlier adjustment with RegArima

Following the notation of Ghysels and Osborn (2001) a seasonal RegARIMA model for a time series with 365 observations per year can be written as

$$\phi_p(B)\phi_P(B^{365})(1-B)^d(1-B^{365})^D\left(y_t - \sum_{i=1}^{r}\beta_i x_{it}\right) = \theta_q(B)\theta_Q(B^{365})\varepsilon_t \qquad (2.10)$$

where $\phi(B)$ and $\theta(B)$ are polynomials of order p, P, q, and Q, while B is the backshift operator. β_i captures the impact of the i-th regressor x_{it} on the time series y_t and ε_t is the error term. The ARIMA part of Equation (2.10) can be written in short form as $(p\,d\,q)_1(P\,D\,Q)_{365}$ with the AR order p, MA order q and number of differences d and capitals indicating seasonal terms (for a thorough discussion of ARIMA modelling see Box and Jenkins, 1970; Brockwell and Davis, 1987).

As a higher periodicity increases the computational burden, it is not always feasible to use seasonal ARIMA models with daily series. In these situations, popular alternatives include estimating the seasonal factor by trigonometric functions or dummy variables in a non-seasonal RegARIMA model. The former can be expressed as a function of sines and cosines (Campbell and Diebold, 2005; De Livera et al., 2011):

$$\hat{S}_t = \sum_{j=1}^{J}\left(\hat{\beta}_{1,j}\, sin\left(\frac{2\pi j G(t)}{365}\right) + \hat{\beta}_{2,j}\, cos\left(\frac{2\pi j G(t)}{365}\right)\right) \qquad (2.11)$$

where $G(t)$ is a unit step function indicating the day of the year, i.e. runs through 1, ..., 365. The optimal size of J can be determined by an information criterion based on the estimated log likelihood, $ln(\hat{L})$, of the RegARIMA model, which is also used to estimate

$\hat{\beta}_{1,j}$ and $\hat{\beta}_{2,j}$. The corrected Akaike information criterion, $AICc = 2J - 2\ln(\hat{L}) + (2J^2 - 2J)/(T - J - 1)$ with time series length T, may be preferable to the Bayesian information criterion, $BIC = J\,ln(T) - 2\ln(\hat{L})$, as the $AICc$ asymptotically selects the model that minimises the mean squared prediction error (Vrieze, 2012). Also, empirical results show that sometimes the BIC selects a value for J that is close or equal to one, so that only basic annual and no monthly movements are estimated, even if intra-monthly seasonality is clearly present. J will usually be no larger than 30 and thus the trigonometric variant is more parsimonious than the dummy variable approach.

The order of the ARIMA model is either specified by the user or can be determined by automatic model detection. For the latter the Hyndman-Khandakar algorithm is implemented (Hyndman and Khandakar, 2008). There, first the order of non-seasonal and seasonal differencing is determined using unit root tests. Then, the ARMA parametrisation is determined by comparing different ARMA orders iteratively using the $AICc$. Of course, other methods could be used as well (e.g. Gómez and Maravall, 2001).

The outlier adjustment is similar to that described by Chen and Liu (1993). The procedure consists of two parts. Part one contains the iterative one-by-one detection of possible outliers at all time points based on the residuals of an ARIMA model, essentially using t-tests for each and every outlier candidate. Part two includes all potential outliers as regressors in a final ARIMA model and then removes all candidates from the set of potential outliers that do not surpass a prespecified threshold for the t-statistics. For monthly time series, these two parts are iterated until convergence. In the case of daily time series, convergence is infrequent and divergence, i.e. an increasing number of outliers found in each iteration, a serious possibility. To avoid this, the threshold for keeping outliers is higher, often a critical value of the t-statistic of 7 or more is used. Additionally, the number of iterations may be chosen to be smaller than for low frequency time series.

If one or several moving holidays have an impact on the original time series, regressors for these holidays can be incorporated into the estimation process at the stage of the RegARIMA model. Often, it can be difficult to decide on an optimal functional form of the regressors, which always will depend on the characteristics of the given time series. Evidently, the impact of other variables could be accounted for by including appropriate regressors into the RegARIMA model.

For data producers carrying out seasonal adjustment on a regular basis, it can become time-consuming to recalculate the complete seasonal adjustment every time a new data point is added to a series. To overcome the necessity of daily re-estimation, forecasted seasonal factors can be used for the adjustment of additional data. We can extend the series adjusted for intra-weekly seasonality using forecasts of length h based on the final RegARIMA model. As the next two steps, i.e. removing intra-monthly and intra-annual seasonality, are based on the series adjusted for intra-weekly and moving holiday effects, the final seasonally adjusted series will include an h-step ahead forecast. Thus, it is straightforward to compute forecasts for the intra-monthly and intra-annual seasonal factors. The forecasts of the intra-weekly seasonal factors have to be obtained independently.

The forecast of the original series will be given by the forecast of the seasonal factor of the weekday, $S_t^{(7)}$, and the forecast of the original series adjusted for weekday effects, $SA_t^{(7)}$:

$$\hat{Y}_{t+h} = \hat{S}_{t+h}^{(7)} + \widehat{SA}_{t+h}^{(7)}. \tag{2.12}$$

As discussed, $\widehat{SA}_{t+h}^{(7)}$ is obtained using the RegARIMA model. $\hat{S}_{t+h}^{(7)}$ is forecasted using Holt's (2004) double exponential smoothing with heterogeneous seasonal effects to extrapolate the seasonal factors for the day of the week. Thereby, the trajectory of the single weekday-specific seasonal factors are taken into account. The level l_t and the slope b_t of $S_t^{(7)}$ are given by:

$$l_{tj} = \alpha_j S_{tj}^{(7)} + (1-\alpha_j)(l_{t-7,j} + b_{t-7,j}) \qquad \forall \ j \in \ (1, ..., 7) \tag{2.13}$$

$$b_{tj} = \mu_j(l_{tj} - l_{t-7,j}) + (1-\mu_j) b_{t-7,j} \qquad \forall \ j \in \ (1, ..., 7). \tag{2.14}$$

with α_j and μ_j estimated from the data. This is equivalent to estimating double exponential smoothing equations for each weekday separately. Equations (2.13) and (2.14) are used to obtain an h-step ahead forecast:

$$\hat{S}_{t+h,j}^{(7)} = l_{tj} + h b_{tj}. \tag{2.15}$$

Adding the forecast of the intra-weekly seasonal factors to the RegARIMA forecasts yields the forecasts for the original series. These can be used to compute implicit seasonal factors given by the difference between the original and the seasonally adjusted time series.

2.3.3 Step III: Adjusting intra-monthly periodic effects with STL

While the length of any year is always the same except for the leap year, the number of days in a month ranges from 28 to 31 with an average of 30.4 days per month. Ignoring the change of the number of days by assuming each month to have 30 days would blend together the effects of different days of the month.

The problem of confused effects can be circumvented by extending each month to 31 days. There are basically two methods of achieving this. Either the course of each month has to be stretched to 31 days by interpolation after warping the time axis, or the deficit days have to be filled up. The characteristics of the intra-monthly seasonality determine which of these options is preferable. If the seasonal pattern is compressed in shorter months, i.e. peaks and troughs have the same magnitude but occur in less time, then stretching is the optimal response. The latter method is preferable if the typical seasonal movement of months with 31 days is cut-off or halted in the other months.

For the extension of the time series, different methods can be utilised. While regression-based techniques have the advantage of possibly incorporating much or all of the information that is contained in the actual series, in the context of time series analysis they can become computationally expensive. Therefore, it seems preferable to use a more parsimonious approach, namely linear interpolation or cubic splines that can be used for stretching or extending the time series to span over 31 days. Per default in the DSA procedure, the Forsythe-Malcolm-Moler algorithm, i.e. cubic splines that ensure that first and second derivatives of the interpolated values of the time series are smooth, is applied to obtain values for the additional data points needed to have 31 days in each month (Forsythe et al., 1977; Press et al., 1992). Then, the intra-monthly seasonal influences are estimated based on the resulting time series, for which $\gamma_{S(31)}$ has to be chosen to fit the characteristics of the intra-monthly seasonal pattern.

2.3.4 Step IV: Adjusting intra-annual periodic effects with STL

As the time series has been extended in step II, the excess days including every 29th of February are removed so that each year contains 365 days. As in step I and III, the intra-annual seasonality is estimated and adjusted for using STL.[2]

2.3.5 Post-adjustment and forecast

As discussed in Section 2.3.2, the forecasts of the original series are calculated as the sum of the ARIMA based forecasts of the original series adjusted for intra-weekly seasonality and the forecasts of the intra-weekly seasonal factors. The main objective of these forecasts is the provision of seasonal factors so that data providers can use a (controlled) current adjustment scheme for the adjustment of their data (Eurostat, 2015, p. 32). This is of practical importance, because it is time-consuming to re-identify the model, the filters, and the regression parameters each time new data are available.

After all seasonal factors have been estimated, the 29th of February is added to the seasonally adjusted series using a spline interpolation. Finally, the effects of the outliers are added to the seasonally adjusted time series.

[2]It is also possible to extend the year to 366 days by seasonal interpolation of the non-existent 29th of February. This approach is preferable if the seasonal effect in the non-interpolated 29th of February values better describe the trajectory of the seasonal component than the neighbouring observations from February 28th and March 1st.

2.3.6 Implementation

The DSA procedure is implemented in the R package {dsa} and is available on CRAN: https://CRAN.R-project.org/package=dsa.

2.4 Application

2.4.1 Adjusting the German currency in circulation

The currency in circulation is calculated as the cumulated differences of cash outflows from and inflows to the Bundesbank. The distribution of banknotes to the public runs primarily through commercial banks. The series' trend is closely related to nominal GDP, but is also marked by periodical influences (Cabrero et al., 2009). Figure 2.1 highlights the different seasonal fluctuations of the time series. Most noticeably, the currency in circulation increases in Germany some weeks before Christmas, peaking just before the end of December. In addition, the series is characterised by a visible increase towards the end of most months as well as towards the end of the week. Finally, an increased demand for banknotes can be detected around moving holidays such as Easter and Pentecost.

For the currency in circulation in Germany data for all denominations from 5 to 500 Euro are available on a daily basis except for the weekend. The time series analysed here contains the small denominations, namely 5, 10, 20 and 50 Euro, starting from 2011. As no new currency is circulated on the weekend it is assumed that the number of banknotes is the same as on the previous Friday. Equally, missing values on holidays are filled up by the last available observation. As the estimation of the impact of moving holidays uses dummy variables, the estimation of the impact of consecutive holidays can lead to different estimates for the single days, especially because they fall on different days of the week and days of the year.

Using the robust STL, the value for $\gamma_{S(7)}$ that controls the stability of the estimated seasonal factor is set to 53. This takes into account that the weekday effect does not change systematically within the year and is particularly stable. To assess the flexibility of the seasonal factors, one strategy is to inspect the SI-ratio, i.e. the combined seasonal-irregular component, together with the seasonal factor of interest. Both should follow a similar trajectory, except that the seasonal-irregular component will fluctuate more (see Cleveland and Terpenning, 1982).

It will often prove difficult to find the optimal functional form of the regressors that capture the influence of moving holidays and which holidays to include. In the case of the currency in circulation, Easter, Ascension, and Pentecost have an effect on the demand for banknotes, mostly related to travelling and holidays. The effects of these holidays are modelled using dummy variables for each of m days before and for each of

n days after Easter Sunday and Ascension, resulting in $m + n + 1$ dummy variables per holiday. Obviously, this approach has its drawbacks. The appropriate values for m and n are either based on the t-statistics of each dummy variable or information criteria. A practical approach is to estimate a RegARIMA model with large values of m and n and delete all non-significant dummy variables.

The number of cosine and sine terms to be included in Equation (2.11) to capture the monthly and annual seasonal movements is also determined by the lowest AICc which in this case leads to the inclusion of 24 x 2 trigonometric terms. This suggests that intra-annual as well as intra-monthly movements have a significant influence. Note that multiples of 12 capture monthly patterns. To model the moving holiday effects $m = 4$ and $n = 1$ for Easter, while only the day before Ascension and Pentecost Monday are included.

Table 2.1: Estimated moving holiday effects for Currency in Circulation in Germany, in million Euro. Based on Time Series only Adjusted for Intra-Weekly Seasonal Effects. A RegARIMA(1,1,1) model with 2x24 trigonometric terms has been estimated.

	Estimate	S.E.		Estimate	S.E.
Good Friday (t-2)	944.0	124.2	Easter Monday	1131.0	175.9
Good Friday (t-1)	1410.7	175.7	Easter Mon. (t+1)	425.2	124.5
Good Friday	1178.8	199.6	Ascension (t-1)	377.9	74.7
Holy Saturday	1077.8	207.0	Pentecost	239.7	86.0
Easter Sunday	996.0	199.7			

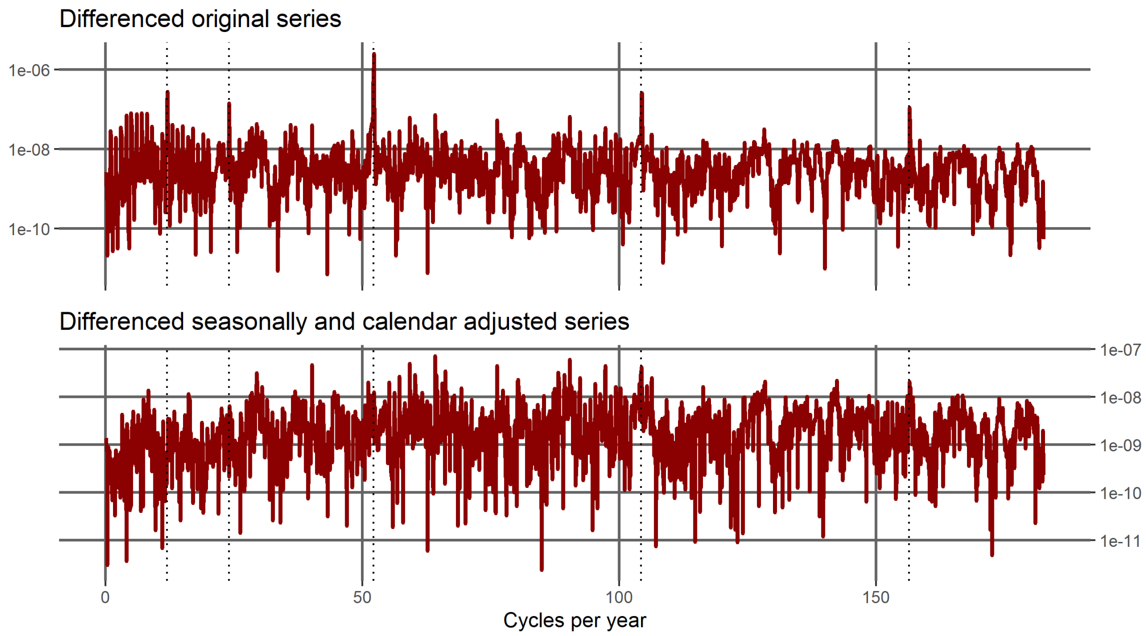

Figure 2.2: Log spectrum of the currency in circulation in Germany

For the computation of the intra-monthly seasonal factors the length of the seasonal Loess regression $\gamma_{S(31)} := 41$ and the robust version of STL is used. Again, this takes into account the high stability of this periodic pattern over time.

For the intra-yearly seasonal factor, the non-robust version of STL is used and the Loess parameter for the seasonal factor $\gamma_{S(365)} := 21$. With only a few years of observations available, the outer loop of STL can sometimes tend to overcorrect seemingly extreme observations. Then it is preferable to use the non-robust version of STL, especially if the seasonal component proves to be stable and a relatively high value of $\gamma_{S(365)}$ is used.

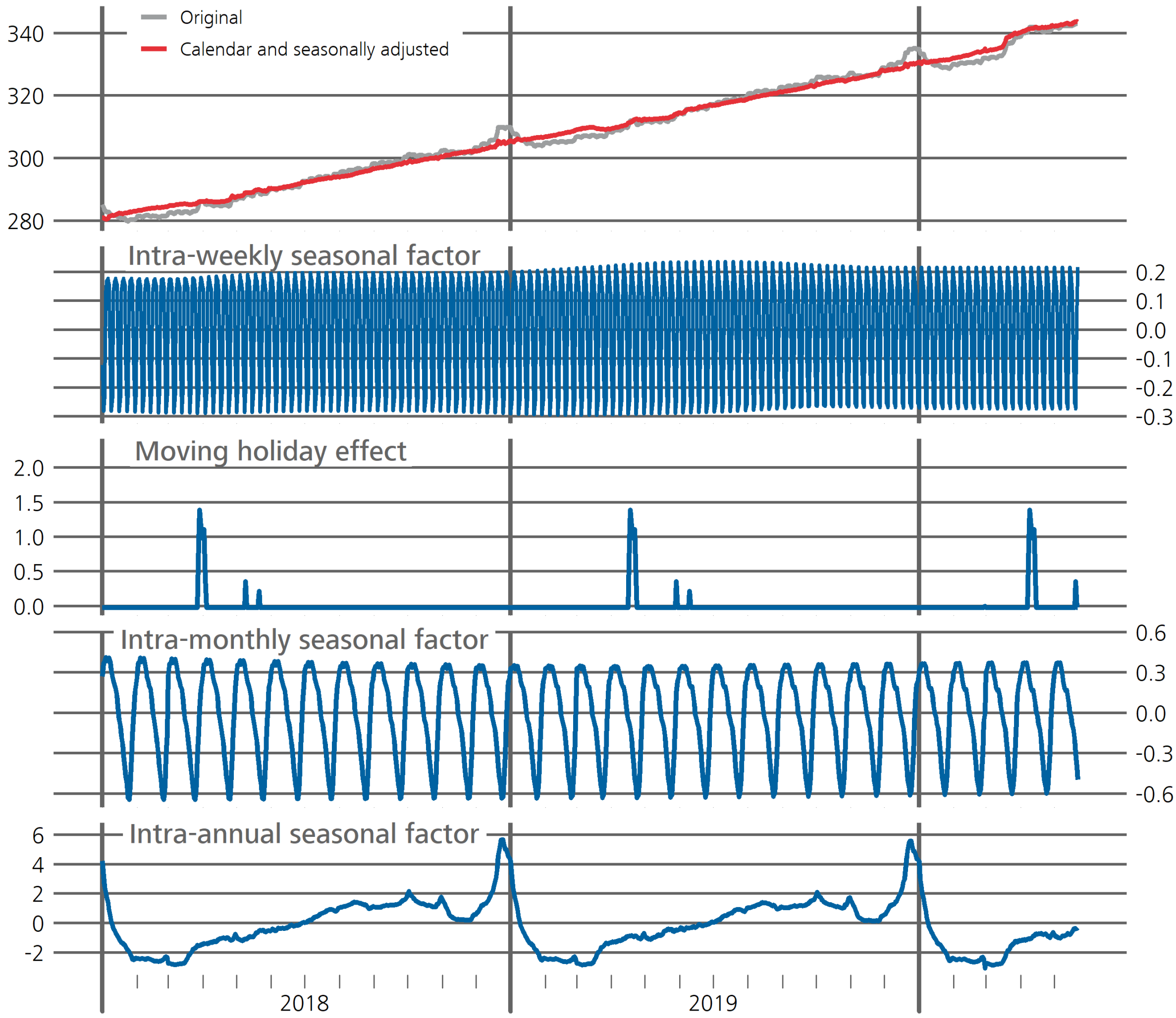

Figure 2.3: Seasonal adjustment results for the currency in circulation in Germany, in billion Euro. Span from Jan 1, 2018 to May 20, 2020.

Comparing the smoothed periodogram (Venables and Ripley, 2002) of the differenced original and the differenced final adjusted series indicates that the spectral peaks at the seasonal frequencies have been filtered out (see figure 2.2). Frequency 12 and 24 are the first monthly cycles; and the first three weekly cycles are indicated at $365/7 \approx 52.1$, $730/7 \approx 104.2$, $1095/7 \approx 156.3$ cycles per year). So based on the spectral diagnostics, we can conclude that the seasonal movements of the currency in circulation have been eliminated successfully. Figure 2.3 shows the original series of the total currency in circulation in Germany and its calendar and seasonally adjusted counterpart.

2.4.2 Adjusting the German electricity consumption

The realised German electricity consumption is compiled by the Bundesnetzagentur [German Federal Network Agency] and is available starting from January, 2015. It has both notable intra-weekly patterns, especially a difference between working days and weekend, and intra-yearly patterns that changes with the seasons (see figure 2.4). Though for most time, the weekday effects are relatively stable, it can change easily in response to economic changes. Therefore, relatively short seasonal filters are used for the day-of-the-week and day-of-the-year effect, while the day-of-the-month is omitted due to missing indications for this effect in the spectrum of the series (see figure 2.5). We use $\gamma_{S^{(7)}} := \gamma_{S^{(365)}} := 13$.

Given that fewer people work on holidays, it is not surprising that there is a lower electricity consumption during holidays. Yet, if a holiday coincides with the weekend, especially Sundays, the decrease is lower compared to holidays falling on working days. This can be modelled explicitly by including weekday specific regressors for each holiday. As we only have less than six years of observations, it can be sensible to combine some of the holidays.[3] Accordingly, all holidays with a fixed date, except those related to Christmas and New Years, are modelled together, weighted by the share of regions for which the day is a public holiday. These are Epiphany (Jan 6), Labour day (May 1), Assumption (Aug 15), German Unity (Oct 3), Reformation Day (Oct 31), and All Saints' Day (Nov 1). In addition to the interaction effects, dummy variables for all moving holidays are added, usually including the day leading up to the holiday and the day after. In the RegARIMA model, 26x2 seasonal trigonometric terms are added to model the seasonality. The seasonal decomposition of the series is shown in figure 2.4.

2.4.3 Adjusting the European NO2 immissions

An important indicator for air pollution is the nitrogen dioxide (NO2) immissions averaged over all available measuring stations in Europe that are made available by the European Environment Agency (EEA). The series is more volatile than the other examples and quite short, starting in January, 2016, so that 4.5 years of observations are available. The series is characterised by a strong and regular day-of-the-week effect and a sinus like annual pattern (see figure 2.6). Balancing the volatility of the series and the dominance of the weekday pattern, $\gamma_{S^{(7)}} := 31$, while $\gamma_{S^{(365)}} := 13$. As there is no indication of day-of-the-month effect, this step in the DSA procedure is omitted. For the estimation of calendar effects, all moving holidays related to Easter that are common in

[3] DSA can be used for time series with at least two years of observations. Yet, the reliability and quality of the decomposition increases with the number of observations available.

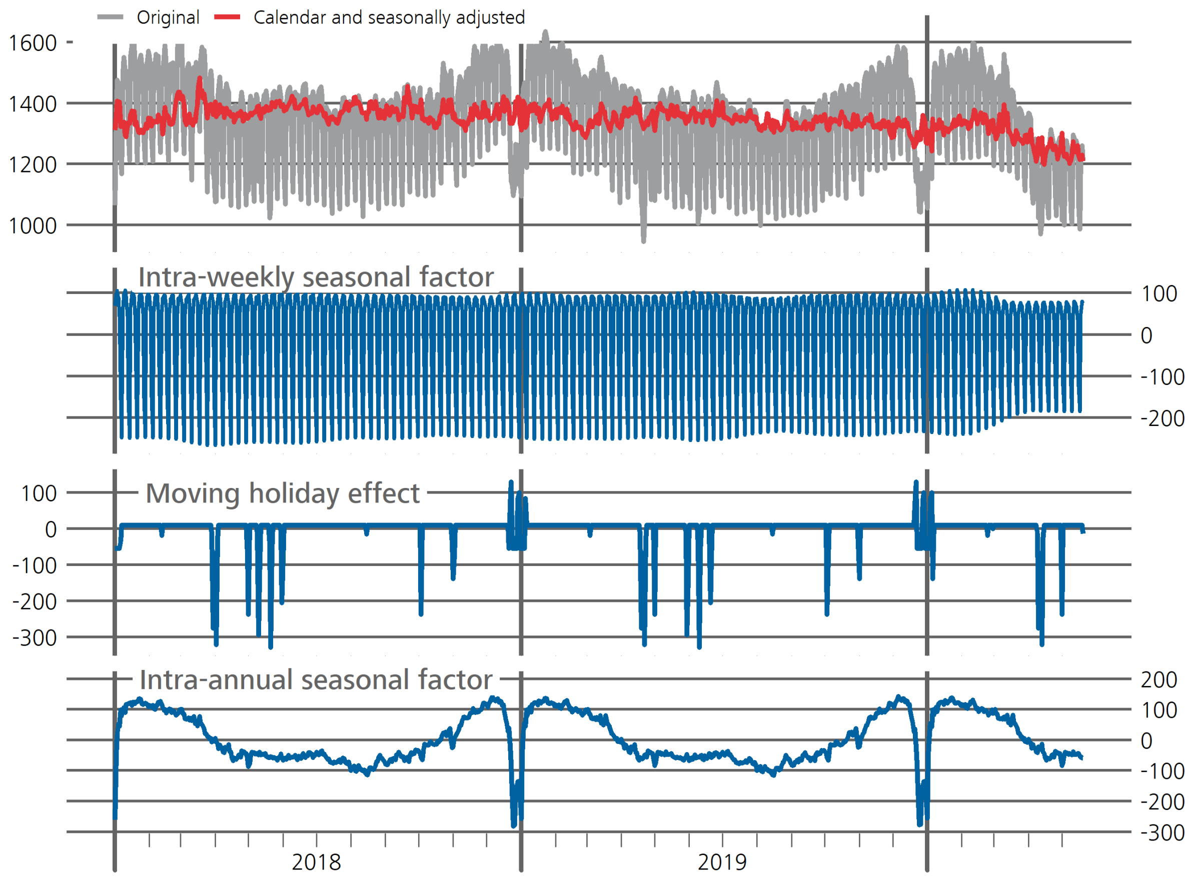

Figure 2.4: Seasonal adjustment results for the German electricity consumption, in GWh.

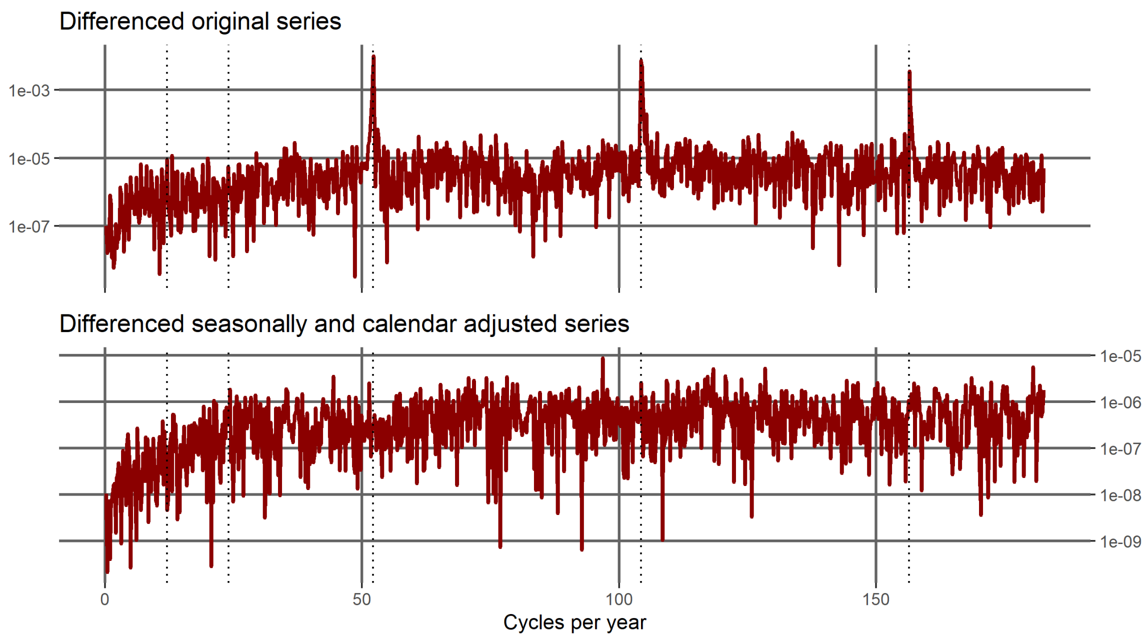

Figure 2.5: Log spectrum for the German electricity consumption, in $\mu g/m^3$. Span from Jan 1, 2018 to May 20, 2020.

Table 2.2: Estimating moving holidays and cross-seasonal effects for the German electricity consumption. Based on Time Series only Adjusted for Intra-Weekly Seasonal Effects. A RegARIMA(313) model with 2x26 trigonometric terms has been estimated.

	Estimate	S.E.		Estimate	S.E.
Carnival Monday	−27.9	8.4	Reformation Day 2017	−226.0	20.9
Holy Thursday	−23.3	10.2	Labour Day (bridge)	−78.3	20.7
Good Friday	−284.5	12.3	National Holidays (Mo-Fr)	−246.8	6.5
Holy Saturday	−82.1	12.7	National Holidays (Sat)	−73.1	17.6
Easter Sunday	−77.2	12.8	German Unity (bridge)	−89.2	15.0
Easter Monday	−330.9	12.3	All Hallows (bridge)	−82.1	14.7
Easter Mon. (t+1)	−38.5	10.2	3d before Xmas (Sat)	56.4	16.3
Ascension (t–1)	−23.5	10.2	Xmas Eve (Sat)	79.3	25.9
Ascension	−302.9	12.7	Dec 26 (Sat)	78.9	23.3
Ascension (t+1)	−145.8	11.1	10d post Dec 26 (Sat)	74.8	12.6
Corpus Christi (t–1)	−19.7	11.2	3d before Xmas (Sun)	120.1	18.9
Corpus Christi	−213.9	13.1	Xmas Eve (Sun)	162.7	23.6
Corpus Chr. (t+1)	−91.9	11.2	Xmas Day (Sun)	175.3	25.9
Pentecost (t–1)	−51.4	11.2	10d post Dec 26 (Sun)	90.9	12.6
Pentecost	−338.3	13.1	Xmas Period (Mo-Fr)	−63.2	10.0
Pentecost (t+1)	−69.5	11.2			

most countries in Europe are considered. This includes Good Friday, Easter Monday, Ascension, Corpus Christi and Pentecost Monday. Easter and Pentecost Sunday do not show significant results, most likely because their impact is indistinguishably similar to that of any other Sunday. Given the simplicity of the annual seasonal pattern, only 1x2 trigonometric terms need to be included in the RegARIMA model. The resulting seasonal decomposition is depicted in figure 2.6. The spectral analysis (see figure 2.7) indicates that periodic patterns have been removed successfully from the unadjusted series.

Table 2.3: Estimating moving holiday effects for European NO2 immissions, in $\mu g/m^3$. Based on Time Series only Adjusted for Intra-Weekly Seasonal Effects. A RegARIMA(112) model with 2x1 trigonometric terms has been estimated.

	Estimate	S.E.		Estimate	S.E.
Good Friday	−2.9	0.6	Corpus Christi	−1.5	0.7
Easter Monday	−5.6	0.6	Pentecost	−3.4	0.7
Ascension	−3.2	0.6			

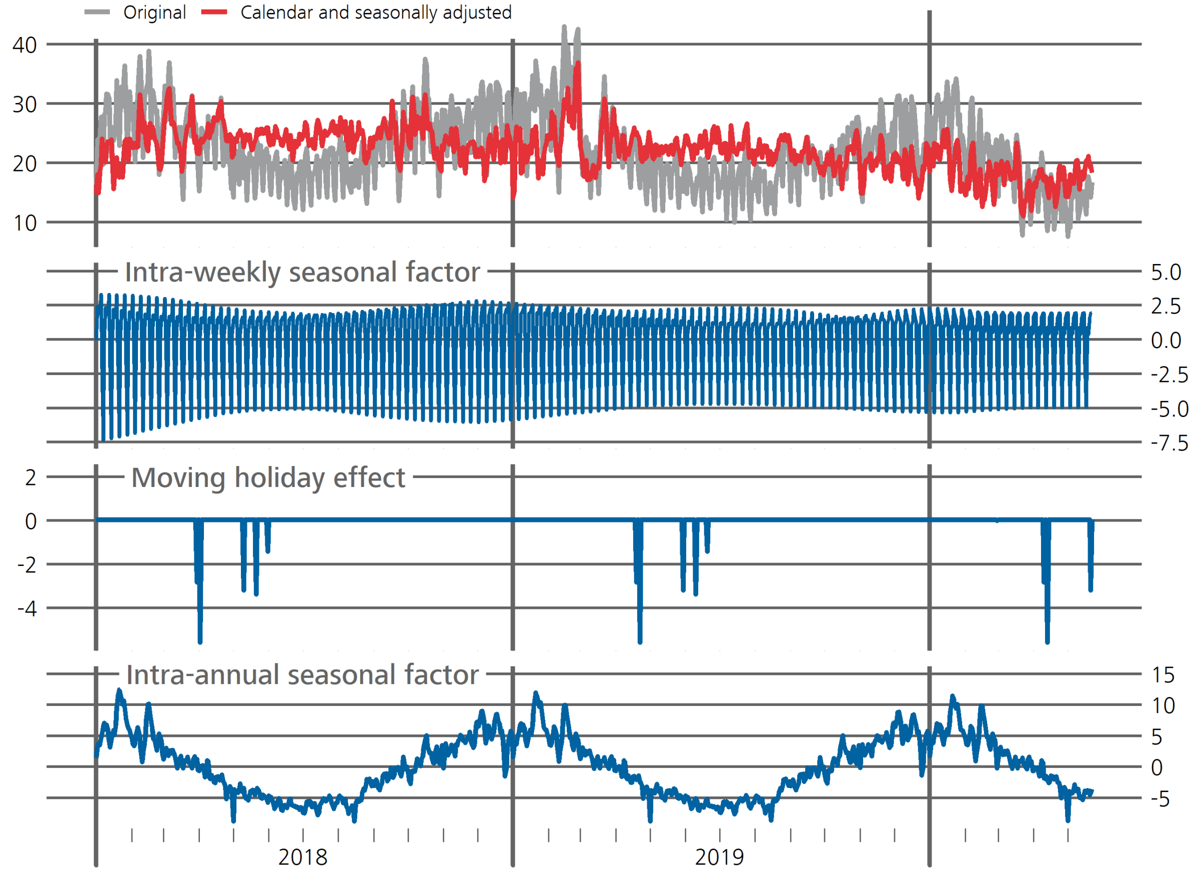

Figure 2.6: Seasonal adjustment results for European NO2 immissions, in $\mu g/m^3$

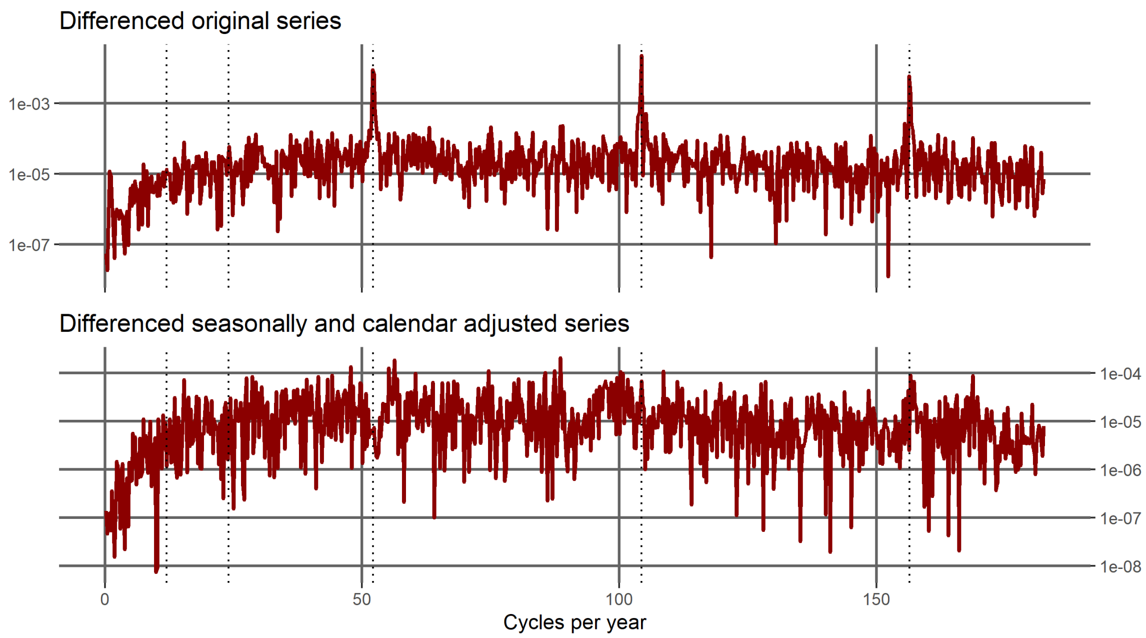

Figure 2.7: Log spectrum of European NO2 immissions, in $\mu g/m^3$

Table 2.4: Comparison of different seasonal adjustment procedures based on seasonality tests for the German currency in circulation.

	Original		DSA		STR		TBATS	
	Stat	P-value	Stat	P-value	Stat	P-value	Stat	P-value
Tests on intra-yearly seasonality								
Friedman	634.8	0.000	147.1	1.000	412.7	0.040	555.7	0.000
QS	0.0	1.000	0.0	1.000	0.0	1.000	10.6	0.005
Tests on series aggregated to months								
Friedman	80.6	0.000	6.5	0.840	16.5	0.125	9.4	0.586
QS	128.1	0.000	4.2	0.125	43.7	0.000	2.8	0.252
Tests on weekday-effects (only last 10 weeks)								
Friedman	16.6	0.011	9.6	0.141	4.2	0.655	5.5	0.478
QS	11.3	0.004	5.4	0.065	6.0	0.049	1.0	0.593
Tests on weekday-effects (whole series)								
Friedman	1,568.2	0.000	28.4	0.000	753.4	0.000	496.7	0.000
QS	681.9	0.000	0.0	1.000	23.6	0.000	0.0	1.000

Table 2.5: Comparison of different seasonal adjustment procedures based on seasonality tests for the German electricity consumption.

	Original		DSA		STR		TBATS	
	Stat	P-value	Stat	P-value	Stat	P-value	Stat	P-value
Tests on intra-yearly seasonality								
Friedman	195.570	1.000	84.6	1.000	404.8	0.069	498.8	0.000
QS	0.0	1.000	0.0	1.000	97.9	0.000	180.4	0.000
Tests on series aggregated to months								
Friedman	48.0	0.000	1.2	1.000	21.8	0.026	4.0	0.971
QS	60.8	0.000	0.0	1.000	4.8	0.092	0.0	1.000
Tests on weekday-effects (only last 10 weeks)								
Friedman	45.8	0.000	8.0	0.237	10.8	0.096	31.1	0.000
QS	55.9	0.000	0.0	1.000	0.8	0.664	7.1	0.029
Tests on weekday-effects (whole series)								
Friedman	1,381.2	0.000	6.4	0.378	303.6	0.000	207.5	0.000
QS	2,772.3	0.000	0.0	1.000	0.0	1.000	1.6	0.453

2.5 Comparison with other procedures

In official statistics there is no established seasonal adjustment routine for daily data. Among the procedures that are capable of handling daily time series, STR (Dokumentov and Hyndman, 2015) and TBATS (De Livera et al., 2011) are available in the R packages {stR} and {forecast} and will be used for the following comparison.

Table 2.6: Comparison of different seasonal adjustment procedures based on seasonality tests for European NO2 immissions.

	Original		DSA		STR		TBATS	
	Stat	P-value	Stat	P-value	Stat	P-value	Stat	P-value
Tests on intra-yearly seasonality								
Friedman	313.7	0.973	70.0	1.000	448.8	0.002	318.4	0.959
QS	0.0	1.000	0.0	1.000	28.1	0.000	0.0	1.000
Tests on series aggregated to months								
Friedman	29.8	0.002	0.4	1.000	0.5	1.000	26.7	0.005
QS	14.4	0.001	0.0	1.000	0.0	1.000	12.7	0.002
Tests on weekday-effects (only last 10 weeks)								
Friedman	23.8	0.001	5.4	0.488	6.8	0.343	21.7	0.001
QS	17.5	0.000	0.0	1.000	0.0	1.000	14.8	0.001
Tests on weekday-effects (whole series)								
Friedman	884.7	0.000	7.4	0.289	2.0	0.917	857.6	0.000
QS	1,397.5	0.000	5.3	0.071	0.0	1.000	1,311.5	0.000

The seasonal adjustment results will be evaluated graphically as well as using tests on (residual) seasonality. The seasonality tests used here are the seasonal autocorrelation based QS test and the ANOVA-type ranks based Friedman test that have been identified by Ollech and Webel (2022) as having a high accuracy in assessing the seasonality of monthly time series.[4] Yet, the QS test is advocated by Ladiray et al. (2018) to be used for daily data. The Friedman test is a sensible addition, as it is particularly good at identifying stable seasonal patterns. These tests will be applied to the original data, and the series adjusted by DSA, STR and TBATS, respectively. STR and TBATS are applied using default values including automatic identification of optimal parameters, the seasonal periods considered are $\tau = \{7, 365.2524/12, 365.2524\}$ for the currency in circulation and $\tau = \{7, 365.2524\}$ for the German electricity consumption and European NO2 immissions. The DSA settings are those described in the previous chapters.

The day-of-the-year effect will be assessed directly and on the series aggregated to months. The day-of-the-month effect will not be tested explicitly, as the test results would depend on the transformations of the time series. Instead it is assumed that any remaining day-of-the-month effect would be picked up by the seasonality tests on the day-of-the-year. The day-of-the-week effect will be tested in the whole series as well as for the last 10 weeks.

Table 2.4 contains both the test statistic as well as the corresponding p-value for the Friedman and QS test for the German currency in circulation. As we have seen before, the unadjusted series shows stable seasonal patterns with a low amplitude together with low volatility. The seasonality tests can identify a day-of-the-year and a day-of-the-

[4] There has yet not been any analysis on the reliability of different seasonality tests for daily data.

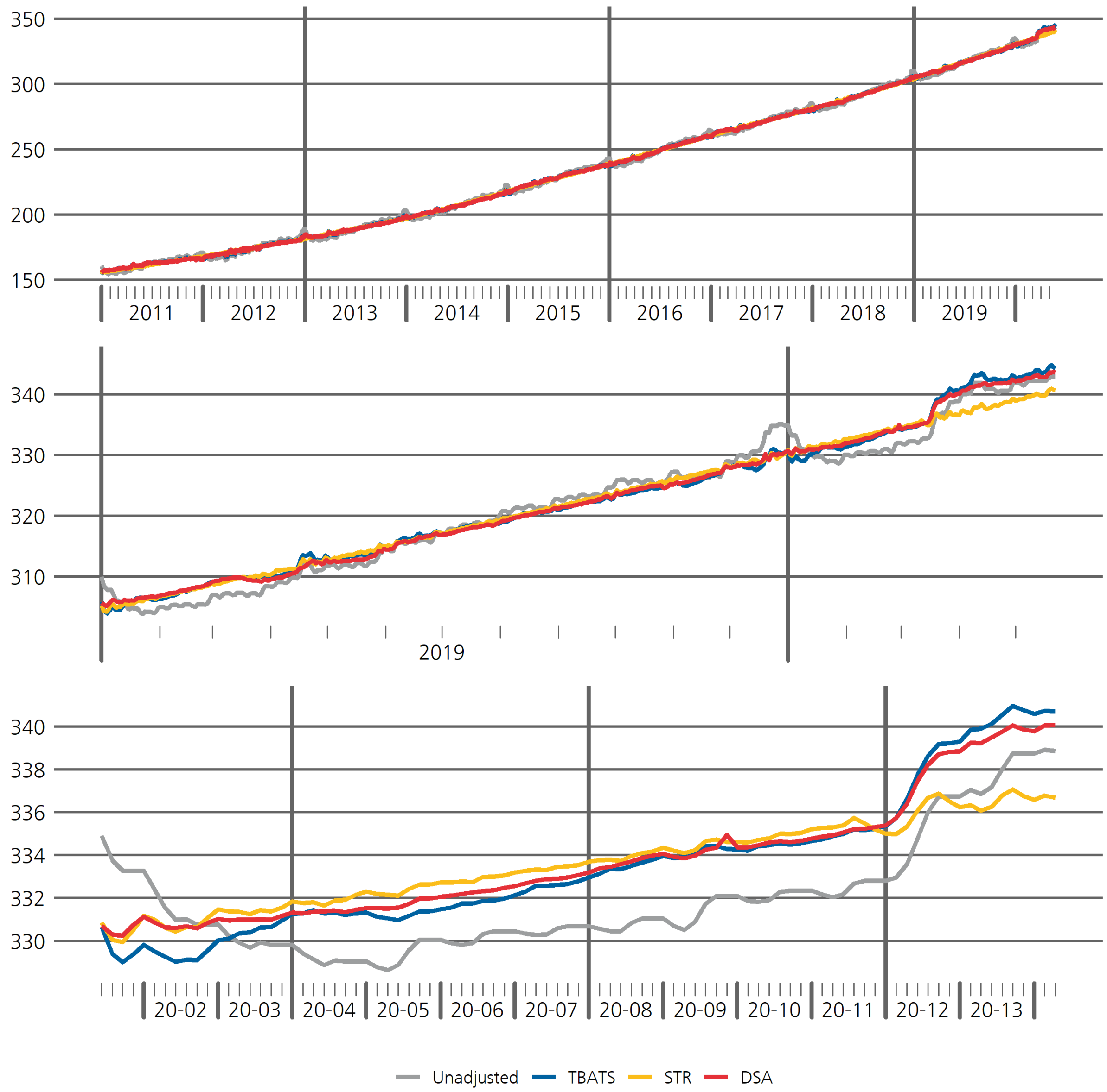

Figure 2.8: Seasonal adjustment results for the Currency in Circulation in Germany, in billion Euro. TBATS and STR have been pre-adjusted for calendar effects using RegARIMA.

week effect in the original series, with the notable exception of the QS test that fails to detect the intra-yearly seasonality in the daily time series. Using the DSA procedure, no residual seasonality can be detected. The significant result for the Friedman test on weekday-effects for the whole series proves to be a consequence of the large number of weeks investigated, as there is no residual weekday-effects in any single year. In general, the results of DSA, TBATS and STR are quite similar for the currency in circulation. Yet, the tests indicate residual seasonality for both intra-yearly and weekday effects and visual inspection of the adjusted series shows that STR does not pick up on a level shift in the trend in the last weeks of the series, i.e. week 12 and 13 in 2020 (see figure 2.8).

For the German electricity consumption, the series adjusted by DSA does not show any signs of residual seasonality. Again, the series adjusted by STR smoothes out a change in the trend, namely, the decrease due to Covid-19 in the Spring of 2020. Both for the daily series as well as the series aggregated to months, the tests partially indicate residual intra-yearly seasonality. The same is true for the series adjusted by TBATS with respect to the intra-weekly pattern, though this is not visible in figure 2.9.

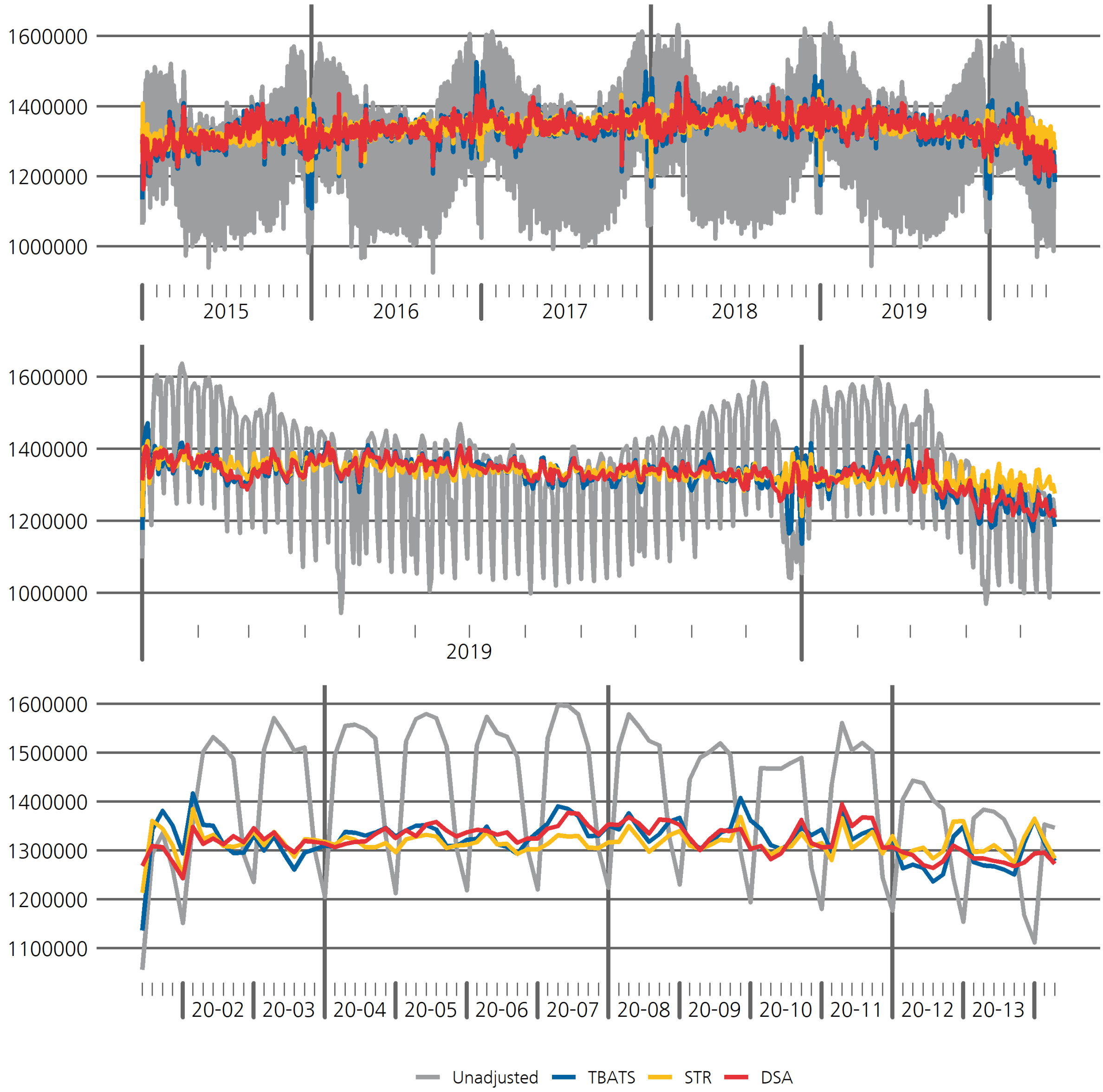

Figure 2.9: Seasonal adjustment results for the German electricity consumption, in GWh. TBATS and STR have been pre-adjusted for calendar effects using RegARIMA.

The European NO2 immissions have strong weekly and annual seasonal patterns, as can be seen in figure 2.10. While both DSA and STR yield similarly seasonally adjusted series - both visibly and informed by the seasonality tests in table 2.6 - TBATS only barely changes the original series and thus remains clearly seasonal.

All in all, only DSA reliably detects and removes the seasonality in all of the time series included here.

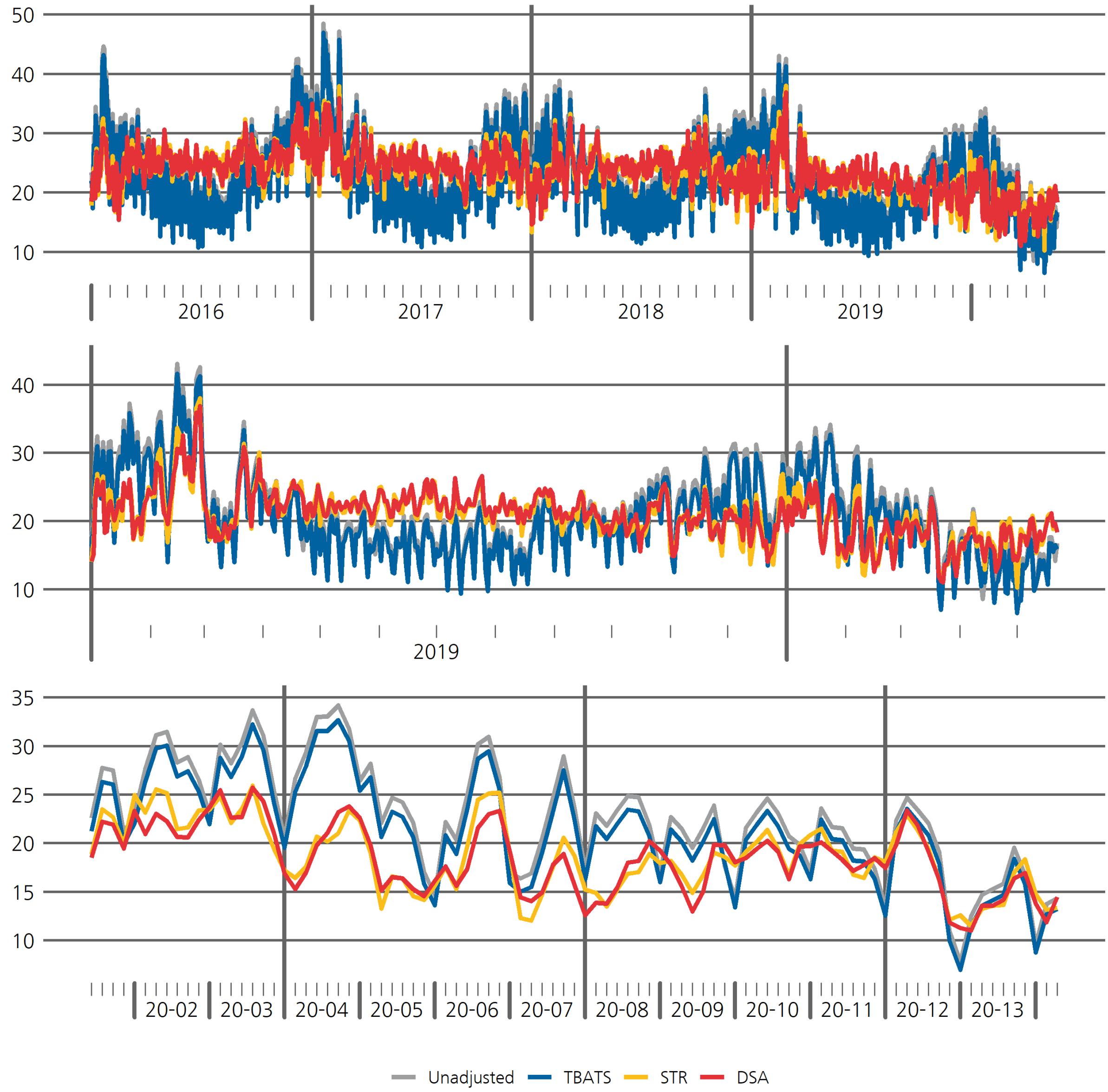

Figure 2.10: Seasonal adjustment results for European NO2 immissions, in $\mu g/m^3$. TBATS and STR have been pre-adjusted for calendar effects using RegARIMA.

2.6 Simulation study

To complement the previous comparison of different seasonal adjustment procedures, we will conduct a simulation study. To gather an understanding of the procedures performance for series with different seasonal intensities, we will vary the magnitude of the seasonal component relative to the non-seasonal part of the series.

To this end, let (pdq) abbreviate the non-seasonal variant of the seasonal ARIMA model given in Equation 2.10, i.e.

$$\phi_p(B)\,\nabla_1^d\,x_t = \theta_q(B)\,\varepsilon_t. \tag{2.16}$$

Using a multiplicative[5] time series model, the simulated daily time series will be given by

$$Y_t = T_t * S_t^{(7)} * S_t^{(365)} * I_t \tag{2.17}$$

including the trend-cycle (T_t), seasonal factors ($S_t^{(\tau)}$) with frequency τ, and an irregular component (I_t). The non-seasonal part of the series, i.e. $T_t * I_t$, will be simulated jointly drawing from a non-seasonal ARIMA model defined in Equation 2.16, so

$$Y_t^{(SA)} = T_t * I_t \sim ARIMA(1, d, 1). \tag{2.18}$$

The seasonal factors are simulated based on a smoothed seasonal random walk. The initial seasonal factor is given by

$$\begin{aligned}\tilde{S}_t^{(\tau)} &= \tilde{S}_{t-\tau}^{(\tau)} + \epsilon_t, \epsilon_t \sim ARIMA(1,1,1),\\ \epsilon_t &= \beta\epsilon_{t-1} + u_t, u_t \sim N(0, \sigma_u^2).\end{aligned}$$

Intuitively, the dependence of ϵ_t on ϵ_{t-1} has the effect that the seasonal factors in adjacent periods are related. σ_u^2 governs the variability of the seasonal component over time. If $\sigma_u^2 = 0$ we have completely stable seasonality; increasing the standard deviation leads to a more dynamic seasonal component.[6]

To ensure that the seasonal factor has the intended magnitude, the standard deviation is fixed. Thus, the seasonal factor is given by

$$S_t^{(\tau)} = \tilde{S}_t^{(\tau)} / \sigma_{\tilde{S}} * \sigma_S$$

with pre-defined σ_S.

The set of simulated daily time series will contain lengths of 4 and 8 years, thus including typically available time series lengths. To ensure that the magnitude of the seasonality depends on the level of the series, a multiplicative relationship between the seasonal and non-seasonal time series components is employed. The seasonal components are designed to cover a broad range of amplitudes. Thus, $\sigma_{S^{(\tau)}} \in \{1, 3, 10\} \;\forall\; \tau \in$

[5] Additive time series models are of course also used in practice, but concerning the simulation, sometimes the level of the series dominates the additive seasonal components and renders them undetectable. This lack of controllability is undesirable here.

[6] For a discussion of stable and dynamic seasonality, see McElroy and Politis (2019, p. 75).

Table 2.7: Root mean square deviation (RMSD) and mean absolute percentage deviation (MAPD) of the seasonally adjusted series by time series lengths

	Total		Time series length			
			4 years		8 years	
	RMSD	MAPD	RMSD	MAPD	RMSD	MAPD
DSA	0.98	0.73	1.00	0.77	0.97	0.72
STR	2.19	1.44	1.42	0.89	2.49	1.72
TBATS	6.75	3.84	6.82	3.82	6.72	3.85

$\{7, 365.2524\}$.[7] The variability of the seasonal components is given by $\sigma_u^2 = 0.01$. $\beta = 0.9$, yielding relatively smooth seasonal factors. As indicated before, the ARIMA models used to obtain the non-seasonal components are of order $(1, d, 1)$, with $d \in \{0, 1\}$, $\phi = 0.5$, and $\theta = 0.8$. Per combination of parameters, 30 series are simulated, yielding a total of 1,080 time series.

For DSA we use the setting from the European NO2 immissions, i.e. $\gamma_{S^{(7)}} := 31$ and $\gamma_{S^{(365)}} := 13$, but with 24x2 trigonometric terms for all series. For STR and TBATS again the default automatic versions are employed, setting the seasonal periods to $\tau = \{7, 365.2524\}$.

For each of the seasonal adjustment procedures considered, table 2.7 shows the deviations of the seasonally adjusted series, $\hat{Y}_t^{(SA)}$, from the simulated non-seasonal series broken down by time series lenghts. The deviations are measured in terms of the root mean square deviation (RMSD) and the mean absolute percentage deviation (MAPD) given by

$$RMSD = \sqrt{T^{-1}\sum_{t=1}^{T}(Y_t^{(SA)} - \hat{Y}_t^{(SA)})^2} \tag{2.19}$$

and

$$MAPD = T^{-1}\sum_{t=1}^{T}\left|\frac{Y_t^{(SA)} - \hat{Y}_t^{(SA)}}{Y_t^{(SA)}}\right|. \tag{2.20}$$

The results are in line with the findings for the real world time series analysed above. For both time series lengths considered, the series adjusted using DSA have on average the lowest deviation from the simulated seasonally adjusted series. The magnitude of the deviations do barely change with time series lengths. Interestingly, this is not the case for STR, for which the deviations almost double for longer time series. TBATS regularly does not capture the seasonal movements and therefore on average deviates considerably from the optimally seasonally adjusted series.

[7] The day-of-the-year effect is simulated with a periodicity of 365 and the leap year is filled up using spline interpolation.

Generally speaking, the deviations from the optimally seasonally adjusted series increase with the size of the simulated seasonal component (see table 2.8). For each magnitude of the seasonal component, DSA outperforms its competitors.

Table 2.8: Root mean square deviation (RMSD) and mean absolute percentage deviation (MAPD) of the seasonally adjusted series by magnitude of seasonal factors

	$\sigma_{S(\tau)=1}$		$\sigma_{S(\tau)=3}$		$\sigma_{S(\tau)=10}$	
	RMSD	MAPD	RMSD	MAPD	RMSD	MAPD
Day-of-the-week component, $\tau = 7$						
DSA	0.96	0.72	0.98	0.73	1.05	0.78
STR	1.72	1.29	1.72	1.28	1.80	1.34
TBATS	2.60	2.00	3.45	2.66	8.90	6.85
Day-of-year component, $\tau = 365.2524$						
DSA	0.79	0.59	0.84	0.64	1.38	1.01
STR	0.98	0.73	1.34	1.00	2.93	2.18
TBATS	3.55	2.75	3.97	3.06	7.44	5.70

2.7 Conclusion

The procedure presented here closes a gap in the methods for seasonal and moving holiday adjustment of daily time series. As shown with a set of diverse daily time series, the procedure yields sensible results and is able to adjust the intra-weekly, intra-monthly and intra-annual periodic movements in addition to moving holidays and cross-seasonal effects. Comparing the seasonal adjustment results with both STR and TBATS shows, that the DSA procedure more reliably identifies and adjusts seasonal patterns from the data.

Future developments should aim at improving the computational speed and reliability of the outlier detection and estimation. Also, automatic default settings for the length of the seasonal Loess regression $\gamma_{S(f)}$ based on the characteristics of a given time series should be investigated. Additionally, post-smoothing of the seasonal factors may increase the stability of the seasonal factors for series with highly harmonic, steady seasonal movements and few observations.

2.A Appendix: Seasonality tests

Given a weakly stationary time series $\{y_t\}$ with length T, both the QS and Friedman test the null hypothesis of no seasonality, i.e. absence of a periodically recurring pattern with cycle length τ.

The QS test does this by testing for significant positive autocorrelation at the first two seasonal lags. Let $\hat{\rho}(h)$ denote the estimated lag-h autocorrelation of $\{y_t\}$. Then, the null hypothesis is specified as $H_0 : \rho(k) \leq 0$ for $k \in \{\tau, 2\tau\}$, and the test statistic is given by

$$QS = T\,(T+2)\left(\frac{\hat{\rho}^2(\tau)}{T-\tau} + \frac{\left[\max\{0, \hat{\rho}(2\tau)\}\right]^2}{T-2\tau}\right)\mathbb{1}_{\{\hat{\rho}(\tau)>0\}}. \tag{2.21}$$

It can be shown by Monte Carlo simulation that the null distribution of the QS-statistic for daily data can also be approximated by a $\chi^2(2)$-distribution.

The Friedman test is a one-way ANOVA with repeated measures (Friedman, 1937) computed on the year-wise ranks of the series. Let r_{ij} be the rank of the observation in the i-th period of the j-th year and μ_i the mean rank over all observations in the i-th period. The null hypothesis is then given by $H_0 : \mu_1 = \mu_2 = \cdots = \mu_\tau$, and the test statistic is defined as

$$FT = \frac{\tau - 1}{\tau}\sum_{i=1}^{\tau}\frac{n\left[\bar{r}_i - (\tau+1)/2\right]^2}{(\tau^2-1)/12}, \tag{2.22}$$

where $\bar{r}_i = 1/n\sum_{j=1}^{n} r_{ij}$. Under H_0, the FT-statistic follows asymptotically a $\chi^2(\tau - 1)$-distribution.

Bibliography

Akaike, H. (1980). Seasonal Adjustment by a Bayesian Modeling. *Journal of Time Series Analysis*, 1(1):1–13.

Bartzsch, N., Seitz, F., and Setzer, R. (2015). The Demand for Euro Banknotes Issued in Germany: Structural Modelling and Forecasting. Technical report, ROME Discussion Paper Series.

Box, G. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Brockwell, P. and Davis, R. (1987). *Time Series: Theory and Methods*. New York: Springer.

Cabrero, A., Camba-Mendez, G., Hirsch, A., and Nieto, F. (2009). Modelling the Daily Banknotes in Circulation in the Context of the Liquidity Management of the European Central Bank. *Journal of Forecasting*, 28(3):194–217.

Campbell, S. D. and Diebold, F. X. (2005). Weather Forecasting for Weather Derivatives. *Journal of the American Statistical Association*, 100(469):6–16.

Chen, C. and Liu, L.-M. (1993). Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*, 88(421):284–297.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6(1):3–73.

Cleveland, W. P., Evans, T., and Scott, S. (2018). Weekly seasonal adjustment: A locally-weighted regression approach. In Mazzi, G. L., Ladiray, D., and Rieser, D. A., editors, *Handbook on Seasonal Adjustment*, chapter 28, pages 757–783. Publications Office of the European Union, Luxembourg.

Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Cleveland, W. S. and Terpenning, I. J. (1982). Graphical Methods for Seasonal Adjustment. *Journal of the American Statistical Association*, 77(377):52–62.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527.

Dokumentov, A. and Hyndman, R. J. (2015). STR: A Seasonal-Trend Decomposition Procedure Based on Regression. Technical report, Monash University, Department of Econometrics and Business Statistics.

Eurostat (2015). *ESS Guidelines on Seasonal Adjustment*.

Forsythe, G. E., Malcolm, M. A., and Moler, C. B. (1977). *Computer Methods for Mathematical Computations*. Prentice-Hall Series in Automatic Computation. Englewood Cliffs, New Jersey. Prentice Hall.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701.

Ghysels, E. and Osborn, D. R. (2001). *The Econometric Analysis of Seasonal Time Series*. Cambridge University Press.

Gómez, V. and Maravall, A. (2001). Automatic Modeling Methods for Univariate Series. In Peña, D., Tiao, G. C., and Tsay, R. S., editors, *A Course in Time Series Analysis*, pages 171–201. Wiley, New York.

Harvey, A., Koopman, S. J., and Riani, M. (1997). The Modeling and Seasonal Adjustment of Weekly Observations. *Journal of Business and Economic Statistics*, 15(3):354–368.

Holt, C. C. (2004). Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. *International Journal of Forecasting*, 20(1):5–10.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(3):1–22.

Koopman, S. J., Harvey, A. C., Doornik, J. A., and Shephard, N. (2009). *Structural Time Series Analyzer, Modeller, and Predictor (STAMP 8.2)*. Timberlake Consultants Ltd.

Koopman, S. J. and Ooms, M. (2003). Time Series Modelling of Daily Tax Revenues. *Statistica Neerlandica*, 57(4):439–469.

Koopman, S. J. and Ooms, M. (2006). Forecasting Daily Time Series Using Periodic Unobserved Components Time Series Models. *Computational Statistics & Data Analysis*, 51(2):885–903.

Ladiray, D., Palate, J., Mazzi, G. L., and Proietti, T. (2018). Seasonal Adjustment of Daily and Weekly Data. In Mazzi, G. L., Ladiray, D., and Rieser, D. A., editors, *Handbook on Seasonal Adjustment*, chapter 29, pages 757–783. Publications Office of the European Union, Luxembourg.

McElroy, T. S. and Politis, D. N. (2019). *Time Series. A First Course with Bootstrap Starter*. Chapman and Hall/CRC, 1 edition.

Ollech, D. (2018). Seasonal adjustment of daily data. *Deutsche Bundesbank, Discussion Paper 41/2018*.

Ollech, D. and Webel, K. (2022). A random forest-based approach to combining and ranking seasonality tests. *Journal of Econometric Methods*.

Pierce, D. A., Grupe, M. R., and Cleveland, W. P. (1984). Seasonal Adjustment of the Weekly Monetary Aggregates: A Model-Based Approach. *Journal of Business and Economic Statistics*, 2(3):260–270.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Tukey, J. W. (1960). A Survey of Sampling from Contaminated Distributions. In Olkin, I., editor, *Contributions to probability and statistics : essays in honor of Harold Hotelling*, pages 448–485. Stanford University Press, Stanford, California.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, 4th edition.

Verbesselt, J., Hyndman, R., Newnham, G., and Culvenor, D. (2010). Detecting Trend and Seasonal Changes in Satellite Image Time Series. *Remote Sensing of Environment*, 114(1):106–115.

Vrieze, S. I. (2012). Model Selection and Psychological Theory: A Discussion of the Differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods*, 17(2):228–243.

Young, M. R. (1996). Robust Seasonal Adjustment by Bayesian Modelling. *Journal of Forecasting*, 15(5):355–367.

CHAPTER 3

Economic analysis using higher frequency time series: Challenges for seasonal adjustment

Daniel Ollech

Abstract

The COVID-19 pandemic has increased the need for timely and granular information to assess the state of the economy in real time. Weekly and daily indices have been constructed using higher frequency data to address this need. Yet the seasonal and calendar adjustment of the underlying time series is challenging. Here, we analyse the features and idiosyncrasies of such time series relevant in the context of seasonal adjustment. Drawing on a set of time series for Germany – namely hourly electricity consumption, the daily truck toll mileage, and weekly Google Trends data – used in many countries to assess economic development during the pandemic, we discuss obstacles, difficulties, and adjustment options. Furthermore, we develop a taxonomy of the central features of seasonal higher frequency time series.

Published as: Ollech, D. (2023). Economic Analysis using Higher-frequency Time Series: Challenges for Seasonal Adjustment. Empirical Economics, 64(3):1375-1398.

Keywords: COVID-19; DSA; calendar adjustment; time series characteristics

3.1 Motivation

During the COVID-19 pandemic, policymakers and economists sought more timely and granular information on the state of the economy. To this end, higher frequency indices that track the economic development of a country have been developed. Most prominently, Lewis et al. (2020a) developed the Weekly Economic Index (WEI) for the United States that combines several weekly indicators such as U.S. railroad traffic, electric utility output or unemployment insurance claims. The WEI tracks the output of the U.S. economy and allows the state of the economy to be monitored on a weekly basis (Lewis et al., 2020b, 2021).

The WEI has sparked the development of a series of similar weekly indicators, especially for European countries (e.g. Delle Monache et al., 2020; Eraslan and Götz, 2020; Fenz and Stix, 2021; Wegmüller et al., 2021). Lourenço and Rua (2021) develop a daily variant (DEI) for the Portuguese economy. Woloszko (2020) constructs country-specific economic indicators for 46 countries based on Google Trends data.

During the last decade, the number of seasonal adjustment methods for higher frequency time series has seen a considerable increase (for a discussion and further references see De Livera et al., 2011; Ladiray et al., 2018; Ollech, 2021; Proietti and Pedregal, 2022; Webel, 2020). Yet, in the construction of the WEI, DEI and some of the other indices, these methods are not employed. Instead, ad-hoc measures like year-on-year rates are used to handle the seasonality of the data. This does not usually take into account calendar effects or the structure of the periodic and seasonal effects such as the varying number of weeks per year.

The higher frequency time series used in the construction of the economic indices often present methodological challenges that differ substantially from those encountered in lower frequency data. In this paper, we therefore contribute to the literature by analysing and seasonally adjusting a set of higher frequency time series that is typical of the kind of data relied upon by economists and business cycle analysts during the COVID-19 pandemic. As the literature on the adjustment and modelling of higher frequency seasonal time series is growing rapidly, the lack of a common understanding of the features, that need to be taken into account when modelling and adjusting these data, becomes evident. Therefore, another key contribution of the analysis presented here is the derivation and systematisation of those characteristics that are relevant in the context of seasonal adjustment.

To this end, Section 3.2 discusses the characteristic features of higher frequency time series from a seasonal adjustment perspective. Section 3.3 presents the seasonal adjustment methods employed in this study, while Section 3.4 analyses and adjusts the daily truck toll mileage, the hourly electricity consumption and a weekly Google Trends series.

These series are used as inputs to the Weekly Activity Index (WAI) devised by Eraslan and Götz (2020). Section 3.5 summarises.

3.2 Key notions for the adjustment of higher frequency data

Seasonal higher frequency time series show features and idiosyncrasies that are of immediate importance for any seasonal adjustment procedure. The time series examples discussed in the next sections include some or all of the characteristics summarised in Table 3.1. To guide the readers understanding and as a point of reference, this table is included at the beginning of this paper.

Table 3.1 organises the characteristics of seasonal higher frequency time series into a taxonomy of relevant features.

The basic characteristics cover the most obvious features that concern the whole set of observations. The fact that higher frequency time series contain many observations – especially in comparison to lower frequency time series – is in itself trivial. Yet, it increases the computational burden and may render some methods inapplicable. Non-isochronicity and non-equidistance will usually either have to be addressed by adapting the methods used or by transforming the data, e.g. by interpolation or time-warping.

The periodic and calendar effects encompass concepts that extend the definition of seasonality and calendar effects in lower frequency time series. As becomes clear from the examples of daily data, higher frequency time series often contain multiple noticeable periodic effects that may be interdependent and autocorrelational. For calendar effects, we may consider including constellations that are not recommended to adjust for in lower frequency time series, such as bridge days. This may be sensible as the higher number of observations allow us to observe and estimate such effects more directly. For weekly series, the impact of bridge days may even be inseparable from the moving holidays, if they always fall in the same week.

Higher frequency time series are generally more volatile than traditional business-cycle indicators. This is mostly a result of the fact that irregular influences tend to partially offset each other over a longer time span and are therefore less pronounced in monthly and quarterly data. Additionally, many of the higher frequency time series encountered are merely a by-product and are not compiled for economic analysis or official statistics and therefore do not adhere to the same data standards. Accordingly, the methods applied to higher frequency time series usually need to be robust against outliers and irregular observations.

Table 3.1: Key notions for the adjustment of higher frequency data.

Notion	Description
Basic characteristics	
Many observations	Higher frequency data contain many observations which can be a challenge for algorithms and users.
Short series	Many series contain few years of observations.
Temporal aggregation	Feasible temporal aggregation of adjusted series depend on data type.
Non-isochronicity	Number of observations per periodic cycle is not always the same for all cycles, e.g. number of weeks per year.
Non-equidistance	For some series the distance between observations varies, e.g. bank-daily, with 0 to 3+ days between observations.
Date and time conventions	Conventions regarding the start of the week or year numbering impact the data structure.
Periodic and calendar effects	
Multiple periodic effects	Daily time series usually contain day-of-the-week and day-of-the-year effects.
Multilevel periodic effects	Seasonal structure may be hierarchical, e.g. a series with hour-of-the-day and day-of-the-week effects.
Uncommon periodic effects	Higher frequency series may contain other periodic effects, such as week-of-the-month effects.
Breaks in periodic effects	Periodic effects may change rapidly, e.g. as a consequence of fundamental changes in the data generating process.
Cross-seasonality	The periodic and calendar effects can be interdependent.
Autocorrelational seasonality	The seasonal impact of consecutive observations may be highly dependent.
Series-specific calendar effects	Calendar effects can be observed more directly, so regressor construction can more easily be tailored to the series.
Uncommon calendar effects	Higher frequency series may contain other calendar effects, such as daylight saving time.
Bridge days	Bridge days may have a traceable effect on the series.
Outliers and missing values	
Missing values	Due to data availability, some series contain (temporarily) missing values, often at the end of the series.
Unreliable data delivery	Data producers do not necessarily have an obligation to deliver data or provide additional information.
Higher volatility	The volatility of the time series usually decreases with a higher temporal aggregation.
Heteroskedasticity	The volatility may change over time and may be seasonal.
Non-traditional outlier patterns	Observed outlier pattern can be different from lower frequency series, e.g. slower rate of decay in a TC outlier.

3.3 Seasonal adjustment methods

The focus of this paper lies on the qualitative assessment of higher frequency time series and to this end, we employ methods for the seasonal adjustment of such series to guide our understanding. Although we discuss the seasonal adjustment results and highlight potential challenges, we do not advocate for any particular method. Instead, the findings shall carve out features of higher frequency time series that need to be addressed regardless of the method applied.

Let $\{Y_t\}$ denote a series of length T with cycle length τ. For monthly series, τ equals the number of observations per year, i.e. 12. For a daily series with only a weekday pattern, the recurring pattern has a cycle length of 7. The basic time series decomposition is then given by

$$Y_t = T_t + S_t + C_t + I_t \tag{3.1}$$

which includes the trend-cycle (T_t), seasonal (S_t), calendar (C_t) and irregular component (I_t). It can easily be adapted to capture a multiplicative relationship between the components by log-transformation of the original time series.[1] For multiple periodic effects with cycle lengths $\tau_1, \tau_2, ...$, the basic time series model can be generalised to

$$Y_t = T_t + \sum_i S_t^{(\tau_i)} + C_t + I_t \tag{3.2}$$

Remark 1. In Equation 3.2 the periodic effects are subsumed as seasonal effects to allow a parsimonious notation. Here, effects related to calendar constellation are not considered to be periodic effects, even though the calendar constellation recurs with a cycle length of 400 years in the Gregorian calendar. However, this recurrence is not exploited in the modelling of the time series and effects arising from these calendar constellations will not usually recur in similar intensity every 400 years.

As we will see, for some series, different seasonal effects are interdependent. Likewise, calender effects and seasonal effects can be related. If all such interactions are relevant, Equation 3.2 needs to be augmented so that

$$Y_t = T_t + \sum_i S_t^{(\tau_i)} + C_t + \sum_i \sum_{j>i} S_t^{(\tau_i)} * S_t^{(\tau_j)} + \sum_i S_t^{\tau_i} * C_t + I_t. \tag{3.3}$$

In line with the methods used for official seasonal adjustment – namely X-13 and Tramo-Seats – the adjustment of the higher frequency time series presented here combines a seasonal adjustment routine with a RegARIMA based pre-adjustment. The latter is a regression model (Reg) with autoregressive integrated moving average (ARIMA) errors

[1]For a thorough discussion of further decomposition models, see U.S. Census Bureau (2016).

and is used here for the estimation and elimination of calendar and some of the interaction effects. The RegARIMA model is given by

$$\phi_p(B)\phi_P(B^\tau)(1-B)^d(1-B^\tau)^D\left(Y_t - \sum_{i=1}^{r}\beta_i X_{it}\right) = \theta_q(B)\theta_Q(B^\tau)\varepsilon_t \tag{3.4}$$

where $\phi(B)$ and $\theta(B)$ are AR and MA polynomials of order p and q, number of differences d, and capitals indicating seasonal terms while B is the backshift operator, i.e. $B(y_t) = y_{t-1}$. The parameter β_i captures the impact of the i-th regressor X_{it} on the time series Y_t and ε_t is the error term. The ARIMA part of Equation 3.4 can be abbreviated by $(p\,d\,q)(P\,D\,Q)_\tau$. Extensions of this model to multiple seasonalities are known and available (e.g. Svetunkov, 2017), but at the time of writing, these are rarely used in seasonal and calendar adjustment.

3.3.1 Seasonal adjustment of daily time series

The iterative daily seasonal adjustment (DSA) procedure described by Ollech (2018; 2021) combines the aforementioned RegARIMA model – that handles C_t and potentially some interaction effects – with STL (Seasonal and Trend decomposition using Loess, Cleveland et al., 1990).

DSA can integrate other seasonal adjustment methods as well, if they are flexible enough to handle daily data. In this regard, Ollech et al. (2021) discuss the flexibilisation of X-11 to handle higher frequency time series.

STL decomposes a time series into T_t, S_t and I_t using a series of Loess regressions and moving averages to separate out the trend and a periodic pattern from the series. Loess regressions are locally weighted linear or polynomial regressions (Cleveland and Devlin, 1988). Each observation is regressed on a pre-defined set containing the γ_τ closest observations. These observations are weighted, where the weight depends negatively on the distance between observations as follows: With a reference observation – i.e. the observation to be regressed on its neighbouring observations – at time t, the weight of the observation at t_i is given by

$$v_i(t) = \left[1 - \left(\frac{|t_i - t|}{\delta_{\gamma_\tau}(t)}\right)^3\right]^3 \tag{3.5}$$

where $\delta_{\gamma_\tau}(t) = |t_i - t_{\gamma_\tau}|$ is the distance between the γ_τ^{th} farthest t_i and t.

In the robust version of STL, after the decomposition of Y_t into a trend, seasonal and irregular component (inner loop) an extreme value downweighing is added (outer loop). Observations with extreme values get a weight ω_t in the local regressions in the next iterations given by

$$\omega_t = \begin{cases} \left(1 - \left[\frac{|I_t|}{6 \cdot median(|\{I_t\}_{t=1}^N|)}\right]^2\right)^2 & \text{if } |I_t| < 6 \cdot median(|\{I_t\}_{t=1}^N|) \\ 0 & \text{else} \end{cases} \tag{3.6}$$

The inner loop iterates through the following steps:

1. Trend adjustment: $Y_t - T_t^{\{k-1\}} = S_t^{\{k\}} + I_t^{\{k\}} \equiv TA_t^{\{k\}}$, with $T_t^{\{0\}} = \mathbf{0}$.
2. Preliminary periodwise smoothing: Each periodwise subseries of $TA_t^{\{k\}}$ is smoothed by Loess to yield a preliminary seasonal factor $S_{pre,t}^{\{k\}}$. The γ_τ has to be specified by the user as there are no default values available (see Cleveland et al., 1990).
3. Smoothing preliminary seasonal component to capture any low-frequency movements $L_t^{\{k\}}$ from $S_{pre,t}^{\{k\}}$.
4. Obtaining seasonal component: $S_t^{\{k\}} = S_{pre,t}^{\{k\}} - L_t^{\{k\}}$.
5. Seasonally adjusting the original time series: $Y_t^{\{k\}} - S_t^{\{k\}} \equiv SA_t^{\{k\}}$.
6. Obtaining trend: Apply a Loess filter to $SA_t^{\{k\}}$ to extract $T_t^{\{k\}}$.

STL only extracts one periodic pattern at a time. Therefore, DSA combines multiple runs of STL with RegARIMA:

- Step I: Adjust intra-weekly seasonality with STL.
- Step II: Calendar- and outlier adjustment with RegARIMA.
- Step III: Adjust intra-monthly seasonality with STL.
- Step IV: Adjust intra-annual seasonality with STL.

In the intermediate steps, the remaining periodic effects are included in the trend-cycle component, which by default in STL captures low-frequency variation that is lower than the considered seasonal frequency.[2]

The resulting final seasonally adjusted series will be adjusted for all seasonal and moving holiday effects considered. The order of the steps in DSA follows the maxim to start with the periodic pattern with the shortest cycle length, i.e. the intra-weekly seasonality. Unlike the methods used for seasonal adjustment of lower frequency time series, the RegARIMA part is not the first step, as it would necessitate to estimate a RegARIMA model with multiple seasonal parts. Future versions of DSA may include such an extension.

For all computations and analyses we use R 4.0.2 (R Core Team, 2021). The seasonal adjustment of daily time series is computed using the {dsa} package in version 1.1.15.

[2]The default settings used to capture the trend-cycle have not been changed in the applications presented here; for a discussion of potential parameter settings, we refer to Cleveland et al. (1990)

3.3.2 Seasonal adjustment of hourly time series

Hourly series are challenging due to the large number of observations included. For computational reasons, it will often not be possible to estimate a RegARIMA model for such series, if many regressors need to be included. Here we adapt the DSA procedure to the case of hourly data as follows:

- Step I: Adjust hour-of-the-week effects with STL.
- Step II: Calendar- and outlier adjustment with RegARIMA based on daily observations.
- Step III: Adjust hour-of-the-month effects with STL.
- Step IV: Adjust hour-of-the-year effects with STL.

As with DSA, single steps of the routine used to seasonally adjust hourly data may be omitted, e.g. only a few series will exhibit an hour-of-the-month effects. Also, the first step may be changed to model hour-of-the-day effects instead or in addition to day-of-the-week effects (see Remark 9).

Remark 2. The transformation from hourly to daily observations in step II only serves to reduce the computational burden. For some series, it may be feasible and preferable to directly estimate the calendar effects in the hourly series - provided an hourly RegARIMA model can be computed. If a distinct moving holiday impact for each hour of a holiday is assumed, the number of regressors may be extremely large. This is aggravated if cross-seasonal effects need to be modelled, e.g. if interactions between fixed holidays, the weekday and potentially the hour are prevalent.

3.3.3 Seasonal adjustment of weekly time series

On average, a year contains a non-integer number of weeks, namely 52.18. Ladiray et al. (2018) describe how, in these cases, autoregressive fractionally integrated moving average (ARFIMA) processes can be exploited that adapt the seasonal differencing part of Equation 3.4 to incorporate fractionally integrated processes. The authors develop a fractional variant of the well-known airline model, i.e. the seasonal ARIMA model of order $(0,1,1)(0,1,1)_\tau$. For a non-integer seasonal period of $\tau = \lfloor \tau \rfloor + \alpha$, with $\alpha \in [0,1]$, the fractional differencing operator $\tilde{\nabla}_\tau$ can be approximated by a first order Taylor series expansion so that

$$\tilde{\nabla}_\tau Y_t \approx Y_t - (1-\alpha)B^{\lfloor \tau \rfloor} Y_t - \alpha B^{\lfloor \tau \rfloor + 1} Y_t \tag{3.7}$$

The fractional airline model can be used for linearisation of a time series analogously to Tramo.[3] We combine this pre-processing with a SEATS-type time series decomposition of the fractional airline model to seasonally adjust weekly time series with $\tau = 52.18$.

3.4 Empirical illustrations

We will present a small set of higher frequency time series that have been of high importance for economic analysis during the COVID-19 pandemic. We will discuss key features of this series and show how these might be accounted for when conducting calendar and seasonal adjustment.

3.4.1 Truck toll mileage index

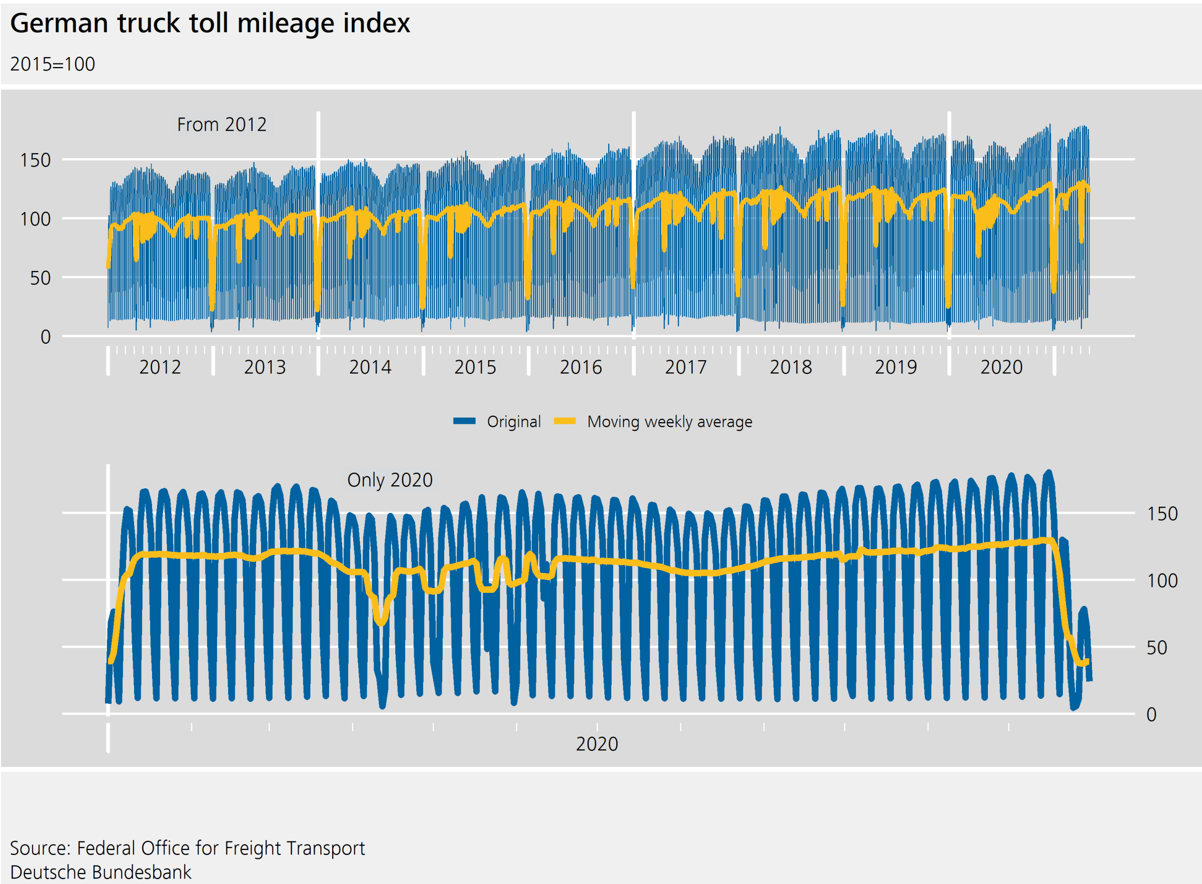

Figure 3.1: German truck toll mileage index

The Federal Office for Freight Transport in Germany is responsible for a distance-based toll on trucks, which was implemented in January 2005. The truck toll mileage

[3]The fractional airline model-based estimation and decomposition are included in the R Package {rjdhf} available on github: https://github.com/palatej/rjdhighfreq. As of version 0.0.4, the Tramo-type estimation of the model does not allow forecasting of the time series. For weekly time series, we therefore use seasonal ARIMA models for pre-processing of the time series with $\tau = 52$.

index has been developed together with the Federal Statistical Office. It is based on raw mileage and free of structural breaks resulting from changes in the vehicles that have to pay the toll. The data are available as a monthly and a daily index (Deutsche Bundesbank, 2020). The daily time series analysed here is available from 1 January 2005 up to 12 September 2021, and thus contains **many observations**.

As can be seen from Figure 3.1 the series is characterised by **multiple periodic effects**, namely a strong weekday pattern, with a trough on Sundays and an annual pattern. These effects are interdependent: the **cross-seasonality** presents itself as a weekday pattern that changes throughout the year. In part, this is due to governing laws and regulations: with only a few exceptions, trucks are not permitted to drive on motorways on Sundays or on public holidays. During July and August, this is extended to Saturdays as well. We further observe different patterns around Christmas related to different changes in consumption behaviour.

Remark 3. Other series contain less typical or even **uncommon periodic effects**. Ollech (2021) finds a monthly recurring pattern in currency in circulation in Germany.

The series is further marked by a weakly positively sloping trend that is halted temporarily in early 2020, as a consequence of the COVID-19 pandemic. After the COVID-19-induced slowdown we observe **breaks in periodic effects**, in particular the weekday pattern. At least temporarily, the difference in the truck mileage between workdays and weekend days is less pronounced. The annual seasonal pattern is **non-isochronous**, i.e. the number of observations per cycle is not the same for all cycles as years contain either 365 or 366 observations.

Remark 4. Non-isochronicity can impact a time series in a number of ways. We may observe that the seasonal pattern in longer cycles is just a stretched-out version of the pattern in shorter cycles. This might be the case, if the seasonal pattern is a smooth pattern in the sense that the seasonal impact of neighbouring values are highly correlated. In series with a more fluctuating pattern, the seasonal impact of additional observations in a given cycle, such as February 29 in a leap year, may be less related to adjacent observations. This may imply that the estimation of the seasonal impact of these additional observations in longer cycles can not be inferred from observations in the shorter cycle.

For higher frequency time series, it is often possible to observe the impact of holidays more directly, and thus identify **series-specific calendar effects**. Figure 3.2 shows the truck toll mileage index on All Saints' Day (November 1) as well as the three days leading up to and the three days after the holiday. All Saints' Day is a public holiday in 5 out of 16 German federal states. In these states, all of which are located either in the south or west of the country, trucks are prohibited from driving on motorways on this day. The impact of All Saints' Day is cross-seasonal: the magnitude of the decrease depends on the

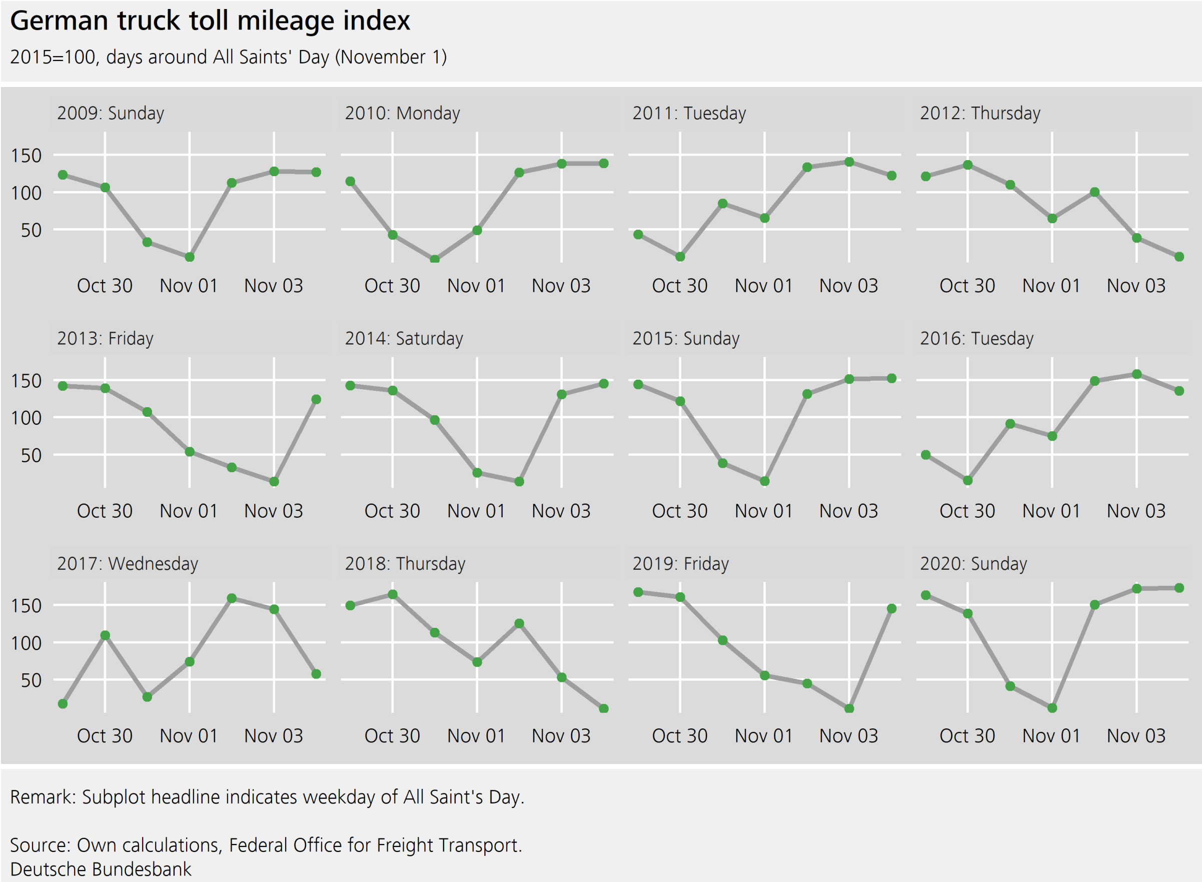

Figure 3.2: German truck toll mileage index

weekday.[4] As restrictions already apply to the driving of trucks on Sundays, All Saints' Day does not have an additional impact on the truck toll mileage index if it falls on a Sunday. By contrast, there is a considerable reduction if it falls on any given weekday.

If All Saints' Day falls on a Tuesday (Thursday), the neighbouring Monday (Friday), i.e. any **bridge days**, also appear to have a reduced number of trucks on motorways. In contrast to lower frequency time series, these bridge day effects can be seen and modelled more directly.

The COVID-19 pandemic, which hit Germany in March 2020, has a twofold impact on the time series. First, it leads to an abrupt decline in the number of kilometres driven by trucks. Second, it impacts the observed weekday pattern owing to changes in the restrictions regarding on which days trucks are allowed to drive on motorways and the consumption pattern.

To obtain a calendar and seasonally adjusted series, DSA was used with $\gamma_7 := 7$ and $\gamma_{365} := 11$ using an additive decomposition. For the seasonal adjustment of the monthly index, we found a multiplicative decomposition to be more appropriate. Yet due to the

[4]Note that Reformation Day - the day before All Saints' Day - is a regional public holiday. Before 2017 it was only a public holiday in all federal states in eastern Germany. In 2017, Reformation Day it was a national public holiday. From 2018, Reformation Day is a public holiday in the federal states in eastern as well as northern Germany.

higher volatility of the daily index – which is typical for higher frequency time series – the seasonal factors from a multiplicative model would inflate strong outliers in many cases, leading to extreme spikes in the seasonally adjusted series. The series does not contain day-of-the-month effects[5]; accordingly, step III of the DSA procedure is omitted. The choice of a very short filter to estimate the day-of-the-week effect, i.e. $\gamma_7 := 7$, allows the day-of-the-week-effect to change during the year, thus making it possible to capture this particular cross-seasonal effect. As discussed above, this is important, because we already know that the weekday pattern is different in different parts of the year, due to the aforementioned driving restrictions. The day-of-the-year effect is captured using $\gamma_{365} := 11$. This is appropriate because it allows the seasonal effect to change slowly over time, but does not overreact to single years. An important tool to identify an appropriate value for γ_τ are visualisations of the seasonal-irregular component (Cleveland and Terpenning, 1982), so-called SI-ratios. These are frequently used to visualise the volatility of the combined seasonal-irregular series and a plausible trajectory of the seasonal component.

To estimate the impact of moving holiday effects and the interaction between weekdays and fixed holidays, a RegARIMA model is used. As described by Ollech (2021), this model combines a non-seasonal ARIMA model with trigonometric terms that capture deterministic seasonality. For the truck toll mileage index, 30 cosine and sine terms are used. Ollech (2021) states that multiples of 12 capture intra-monthly pattern, yet here, the high number of trigonometric terms used instead reflects the complexity of the seasonal pattern and does not indicate a day-of-the-month effect.

Remark 5. To obtain the unadjusted monthly truck toll mileage index, the daily raw truck toll mileage is summed up for each month and transformed into an index. If a **temporal aggregation** is likewise performed on the seasonally adjusted truck toll mileage index via summation, the resulting time series will contain a length-of-the-month effect – i.e. months with more days will have a higher value – and thus be seasonal. Using the average monthly index instead is a simple remedy, but reduces the comparability of the directly adjusted monthly and daily time series.

The calendar and seasonally adjusted truck toll mileage index is especially volatile around holidays (see Figure 3.3). This **heteroskedasticity** is due to estimation uncertainty and the fact that truck drivers are restricted by laws with regards to the number of hours that they are allowed to drive per week and per day. The latter determines the optimal logistics and thus the interaction between holidays, consumer demands and kilometres driven by trucks on a given day resulting in the observed local volatility increases.

Table 3.2 shows the estimated impact of moving holidays and cross-seasonal effects. All moving holidays that are public holidays at the national or federal state level are

[5] The spectrum of the series (not shown here) indicates that the series is not influenced by any monthly recurring periodic pattern. Alternatively seasonality tests may be used to check whether there are significant day-of-the-month effects.

Table 3.2: Estimated moving holiday and cross-seasonal effects for the German truck toll mileage index. Based on time series only adjusted for intra-weekly seasonal effects. A RegARIMA(2,1,1) model with 2x30 trigonometric terms has been estimated.

	Estimate	S.E.		Estimate	S.E.
Carnival Monday	−8.6	1.6	NH[1] (Mon)	−92.5	2.3
Mardi Gras	−5.0	1.6	NH[1] (Tue)	−106.3	2.4
Holy Thursday	−20.2	1.5	NH[1] (Wed, to 2015)	−90.4	3.1
Good Friday	−99.7	1.9	NH[1] (Wed, from 2015)	−118.8	2.7
Holy Saturday	−17.3	1.9	NH[1] (Thu)	−106.7	2.1
Easter Sunday	−10.4	1.9	NH[1] (Fri)	−74.6	1.9
Easter Monday	−117.6	1.9	NH[1] (Sat)	−20.6	1.9
Easter Monday (t+1)	−21.4	1.5	Christmas Period (Mon)	−32.7	1.6
Ascension (t–1)	−13.4	1.5	Christmas Period (Tue)	−42.5	1.7
Ascension (to 2015)	−109.4	2.2	Christmas Period (Wed)	−41.7	1.7
Ascension (from 2016)	−119.4	2.3	Christmas Period (Thu)	−41.4	1.7
Ascension (t+1)	−18.4	1.6	Christmas Period (Fri)	−28.7	1.7
Corpus Christi (t–1, to 2015)	−7.4	2.0	3d before Christmas (Sun)	22.8	2.1
Corpus Christi (to 2015)	−73.8	2.2	Christmas Eve (Sat)	11.7	4.3
Corpus Christi (from 2016)	−74.5	2.1	Christmas Eve (Sun)	42.0	4.8
Corpus Christi (t+1)	−7.7	1.5	Christmas Day (Sat)	24.9	5.9
Pentecost (t-1)	4.1	1.3	Christmas Day (Sun)	58.4	4.3
Pentecost (to 2015)	−107.7	1.9	Boxing Day (Sat)	18.7	3.2
Pentecost (from 2016)	−117.6	2.1	Boxing Day (Sun)	30.4	5.8
Pentecost (t+1)	−14.2	1.5	10d after Dec 26 (Sat)	14.8	1.8
Labour Day (bridge)	−17.9	2.6	10d after Dec 26 (Sun)	30.9	1.9
German Unity (bridge)	−2.6	2.9			
All Saints' Day (bridge)	−20.6	2.4			

[1] *NH includes the following holidays with fixed dates: Epiphany, Labour Day, Assumption Day, German Unity, Reformation Day and All Saints' Day. The weights of the regional holidays are given by: Epiphany 0.2, Assumption Day 0.1, Reformation Day (after 2017) 0.2 and All Saints' Day 0.6.*

Note: Ascension and Labour Day 2008 both fell on 1 May. Because the effect is not additive, the effect has been assigned to Labour Day only, i.e. the regressor for Ascension is 0 *on that day.*

included in the regression. This is extended to the surrounding days, as a transition phase can usually be observed, so that both the day before and after a given public holiday show a decrease. The only exception is the Sunday just before Pentecost Monday, which is slightly positive compared to a typical Sunday in spring. Additionally, we include the main event days of the carnival season for which a significant impact can be detected.

The impact of fixed holidays is usually captured by STL in the last step of DSA. Yet, as discussed above, the interdependency between that impact and the weekday cannot be captured by STL. Here, we model it using interaction dummies in the RegARIMA model. The impact of each weekday is combined across all national and federal state-level holidays to reduce the number of included regressors. Days that are public holidays only regionally are weighted accordingly (for details on the weights see Table 3.2 and Deutsche

Bundesbank, 2020). The remaining spikes around Christmas are not seasonal, as they do not fall on the same day of each year. Yet, with more observations it may be possible to further improve the estimation of cross-seasonal effects for that time period.

The changes in the weekday pattern after the start of the COVID-19 pandemic in Germany discussed above recur weekly and may thus be considered to be part of the $S^{(7)}$-component. A different view is that as these changes are only temporary and a result of the irregular nature of the crisis they should be captured in the irregular component and thus be visible in the calendar and seasonally adjusted series.[6] For the period from 23 March to 30 August 2020, i.e. from the beginning of the lockdown until after the summer holidays, we chose the latter approach. To seasonally adjust this period, forecasted calendar and seasonal factors are used, which are obtained by restricting the estimation span from the beginning of the time series to 22 March 2020. For the period after 30 August, we again use all available data in a controlled current adjustment scheme. This means that the seasonal and calendar components are re-estimated monthly, but we control weekly whether a re-estimation is necessary. This is the case when the forecasted seasonal or calendar component no longer adequately capture the respective effect, e.g. because the seasonal pattern has evolved differently than predicted.

Remark 6. Generally, at the beginning of a crisis, it may be a good policy not to re-estimate the seasonal and calendar components immediately, in order to avoid including transitory and irregular influences into the seasonal or calendar components. Once the crisis has stabilised, a controlled current adjustment scheme as discussed above may be implemented.

3.4.2 Electricity consumption

The German electricity consumption is compiled by the German Federal Network Agency using data from the network providers starting in January 2015. The series analysed here ends on 30 April 2021. Energy consumption is related to industrial production and GDP. Electricity consumption is therefore of particular interest to business-cycle analysis (Arora and Shi, 2016; Do et al., 2016). For the most recent observations, the series is subject to **unreliable data delivery** as some network providers do not always provide data immediately.

Remark 7. More broadly, unreliable data delivery stems from the issue that data providers are often not legally or contractually obligated to provide data and do not

[6] At the time of writing the ESS Guidelines on Seasonal Adjustment does not discuss this matter and therefore does not stipulate any action (Eurostat, 2015). It could be sensible to define a minimum number of cycles that a new periodic pattern need to pass through before it is regarded as a periodic pattern that needs to be adjusted for.

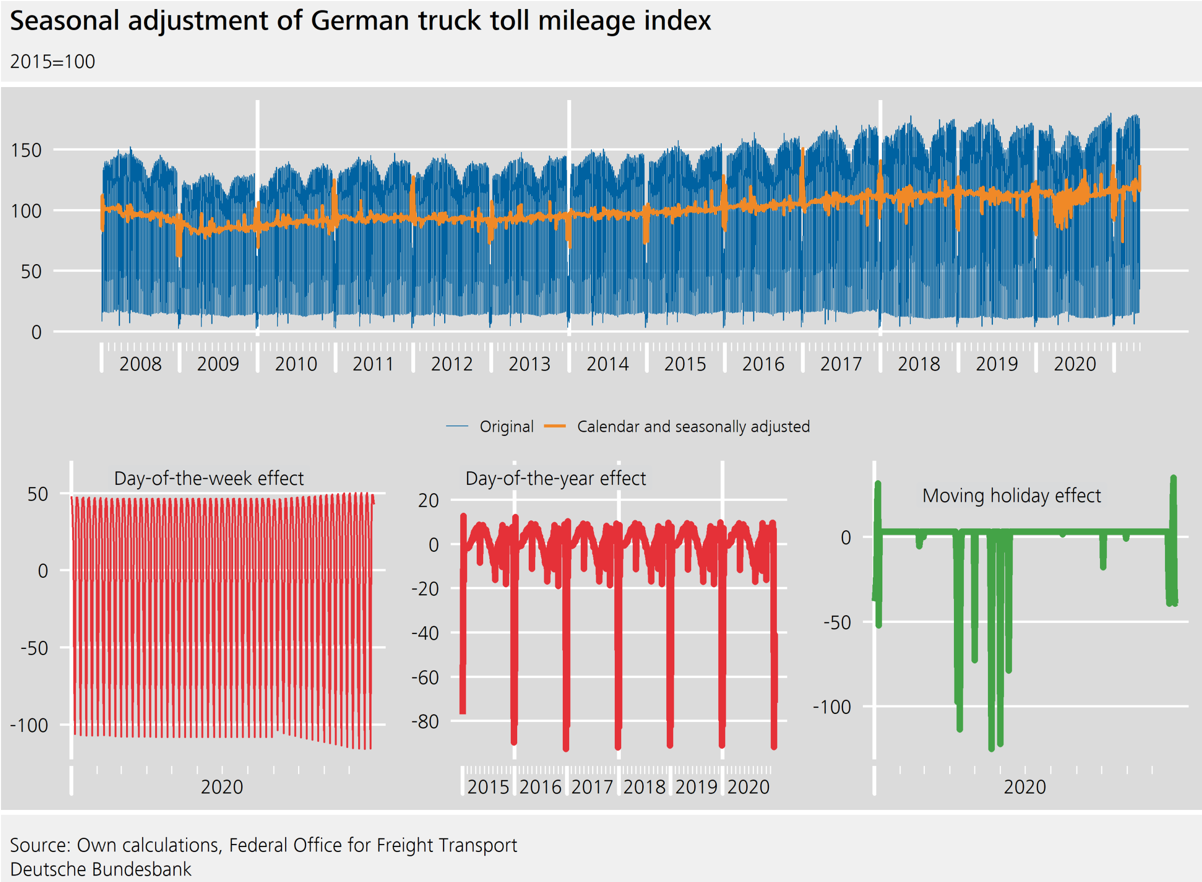

Figure 3.3: Seasonal adjustment of German truck toll mileage index

need to adhere to any standards regarding data quality or deadlines. Furthermore, if the data are merely a by-product, it may be difficult to analyse the quality of the data.

In the case of electricity consumption, this leads to temporarily **missing values** that might require interpolation, either before or as part of the seasonal and calendar adjustment.

Remark 8. Some higher frequency time series do have structurally missing values, e.g. time series that only contain observations on working days. If this is the case, the data may be **non-equidistant**, i.e. the distance in time between observations is not the same for all neighbouring observations. In other words, the distance between a Monday and a Tuesday is less than between a Friday and a Monday.

Electricity consumption is available on a 15-minute basis. Ollech (2021) discusses how the daily electricity consumption can be seasonally adjusted. Here, we will analyse the hourly electricity consumption. Clearly, for business cycle analysis, it may be advantageous to focus on a lower frequency aggregation level, such as a weekly series, as the volatility of the time series tends to decrease with aggregation level. However, discussing an hourly series will inform our understanding of typical characteristics of higher frequency time series and thus is relevant for the modelling of daily and weekly series.

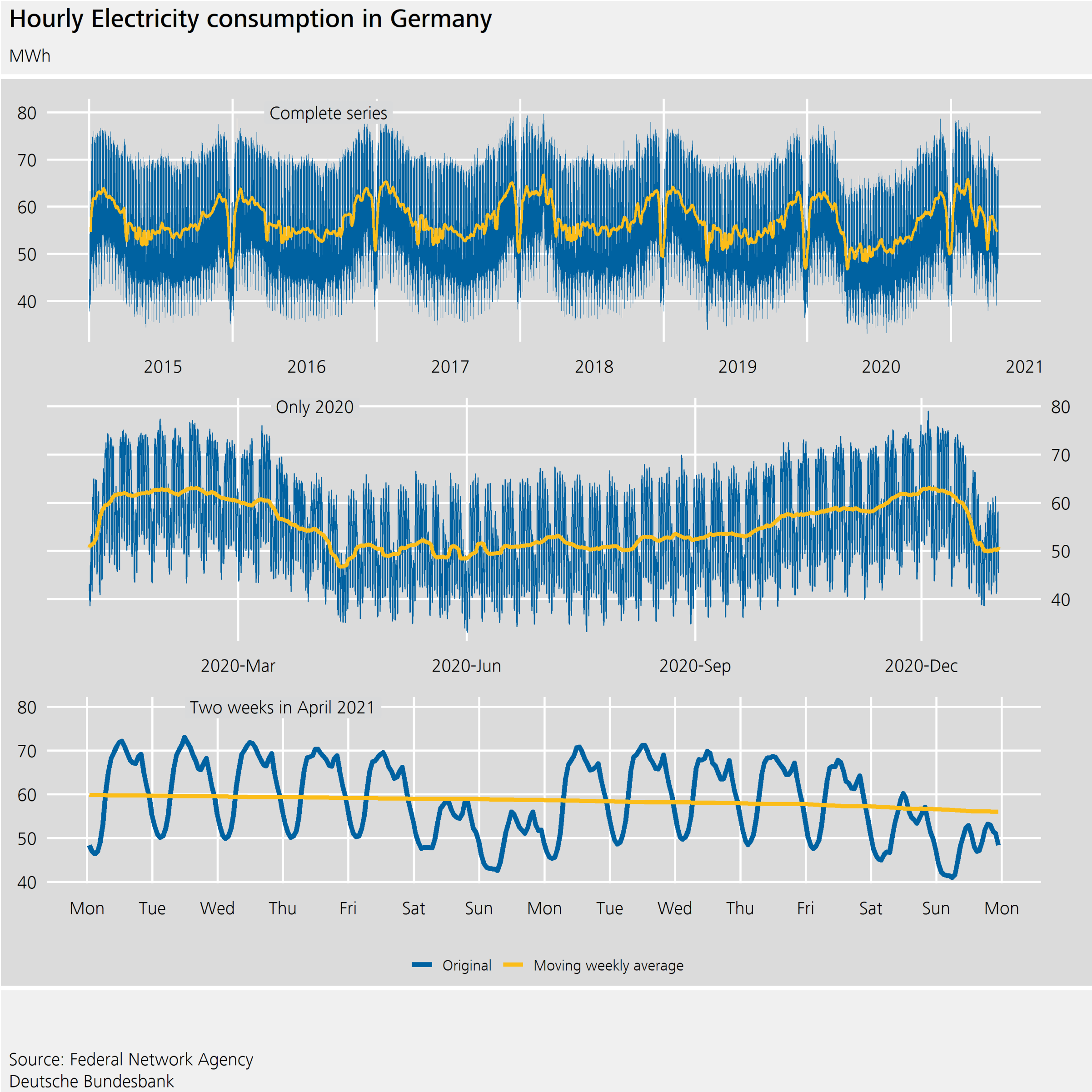

Figure 3.4: Hourly electricity consumption in Germany

As can be seen from Figure 3.4, the electricity consumption is characterised by an hour-of-the-week pattern and – especially if we disregard Christmas and fixed holidays – an almost sinusoidal annual seasonality. The latter is termed **autocorrelational seasonality**, as the seasonal impact of each day of the year strongly correlates with the seasonal influences visible in the neighbouring observations. A dependence structure such as this is usually not exploited fully in the seasonal adjustment of lower frequency time series. Yet, for higher frequency time series it may improve the estimation of the seasonal effects given the high volatility of the time series and the corresponding estimation uncertainty.

Remark 9. Some series show **multilevel periodic effects**. Instead of an hour-of-the-week effect, an hourly series may contain an hour-of-the-day effect and an influence of the day of the week, if the pattern of the intraday-movements remains the same throughout the week and only the level of the series varies from weekday to weekday. By contrast, the magnitude of the intra-day-movements of the electricity consumption changes throughout the week, particularly when contrasting working days and weekends.

The difficulty of estimating all relevant effects is aggravated, as the daily electricity consumption is a relatively **short series**.[7] As a consequence, we cannot observe all possible calendar and cross-seasonal constellations. To illustrate this point, let us assume that we want to model the interaction between the weekdays and Labour Day. With just over five years of observations, not all possible interactions occurred in the time span considered, and in any case, only very few observations per constellation are included.

At times, higher frequency time series show rather **uncommon calendar effects**. The daily electricity consumption is impacted by Daylight Saving Time (DST) – even in the hourly series.

The start of the COVID-19 pandemic in Germany again evokes a temporarily declining trend and some slight and gradual changes in the weekly pattern, reflecting the reduction in output in the production sector and possibly changes in the share of people working remotely (for a cross-country comparison of the impact of the pandemic on electricity consumption, see López Prol and O, 2020). For the seasonal adjustment of lower frequency time series, the pandemic is treated using series of level shifts (LS), additive outliers (AO) and, less frequently, temporary change outliers (TC) with a pre-specified decay rate (Eurostat, 2020). The downturn observed for electricity consumption is gradual enough so that one or multiple level shifts are not necessary to model the crisis based on daily data. In general, we observe **non-traditional outlier pattern** more frequently for higher frequency time series than for monthly or quarterly series. This is neatly illustrated by the finding that the decay rate of 0.7 for TC – the default for monthly series (cf. U.S. Census Bureau, 2016) – is often unsuitable for daily or weekly series.

To estimate the hour-of-the-week effect – which encompasses $24 \cdot 7 = 168$ hours a week – STL is employed with $\gamma_{168} := 7$, i.e. a very short filter that is tailored towards an effect that changes throughout the year. More precisely, the difference between the daily minimum and daily maximum consumption is smaller on the weekend than on working days. This difference is especially small in winter compared to the rest of the year.[8]

After the hour-of-the-week effect has been estimated, the adjusted series aggregated to daily observations serves as input to DSA. We omit the estimation of the day-of-week and day-of-the-month effect, as these effects are not (or no longer) present in this partially adjusted series.

[7] For very short series, it may not be advisable to estimate all seasonal and periodic components. An estimation of the day-of-the-year effect requires at least two years, but for a reliable and sensible estimation usually three or more years of observations are needed. The day-of-the-week effect may often be estimated using less than two years of data.

[8] In daily series this translates into a difference between working days and weekends which increases during the warmer season.

Table 3.3: Estimated moving holiday and cross-seasonal effects for German electricity consumption, in percent. Based on time series adjusted only for hour-of-the week effects and aggregated to a daily series using daily means. A RegARIMA(5,1,1) model with 2x24 trigonometric terms has been estimated.

	Estimate	S.E.		Estimate	S.E.
Carnival Monday	−1.5	0.6	Labour Day (bridge)	−4.8	1.4
Holy Thursday	−1.9	0.7	German Unity (bridge)	−7.0	1.0
Good Friday	−21.6	0.9	All Saints' (bridge)	−4.6	1.0
Holy Saturday	−7.4	0.9	Reformation Day (bridge, after 2017)	−6.8	1.5
Easter Sunday	−6.3	0.9	National Holidays (Mon-Fri)	−17.9	0.4
Easter Monday	−24.2	0.9	National Holidays (Sat)	−4.7	0.9
Easter Monday (t+1)	−3.2	0.7	3d before Christmas (Sat)	1.9	1.1
Ascension (t–1)	−1.6	0.7	Christmas Eve (Sat)	4.3	1.8
Ascension	−22.2	0.9	Dec 26 (Sat)	5.6	1.2
Ascension (t+1)	−9.9	0.7	10d post Dec 26 (Sat)	3.4	0.8
Corpus Christi (t–1)	−1.1	0.7	3d before Christmas (Sun)	6.2	1.3
Corpus Christi	−14.2	0.9	Christmas Eve (Sun)	8.9	1.6
Corpus Christi (t+1)	−6.0	0.7	Christmas Day (Sun)	10.1	1.7
Pentecost (t–1)	1.8	0.6	10d after Dec 26 (Sun)	4.1	0.8
Pentecost	−23.3	0.7	Christmas period (Mon-Fri)	−5.1	0.7
Pentecost (t+1)	−3.1	0.7	Daylight Saving Time Spring	0.5	0.5
			Daylight Saving Time Autumn	−0.9	0.6

Table 3.3 shows the estimated moving holiday and cross-seasonal effects for this time series. The cross-seasonal effects included are again interactions between weekday and fixed holidays. As mentioned, due to the length of the series, not all possible calendar interactions can be observed and overall, only very few observations per constellation are available.

As discussed above, a noteworthy calendar effect is Daylight Saving Time. DST has an obvious – albeit small – effect on the time series. Here, the moving holiday and cross-seasonal effects will be estimated in the series aggregated to a daily series by averaging the hourly observations. Alternatively, the hourly series could be temporally aggregated to a daily series by summing up the hourly values. Generally, if a multiplicative time series model is used, the difference between the estimated effects using daily averages comparing to those obtained using daily sums is often negligible for calendar constellations that impact the whole day. For example, electricity consumption is estimated to be 23.3 percent lower on Pentecost based on daily averages (see Table 3.3). If we used daily sums instead, the estimated impact were 23.2 percent.

If we consider matters that affect only single hours of a day, such as DST, the differences can be considerable. Transforming the hourly to a daily series by taking the mean will reduce the estimated impact of DST on the series, because DST will mostly be reduced to capturing the configuration of the hours in a day and the effect on the electricity consumers. If we used hourly sums instead, the estimated impact of DST would additionally include a length-of-the-day effect.

After the calendar and cross-seasonal effects have been estimated, they are broken down into hours, assuming that all hours of the day are influenced in the same way. The hourly series is then adjusted using these hourly factors.

Finally, the hour-of-the-year effect is estimated using STL in DSA with $\gamma_{24 \cdot 365} := 13$. The final adjusted series can be seen in Figure 3.5.

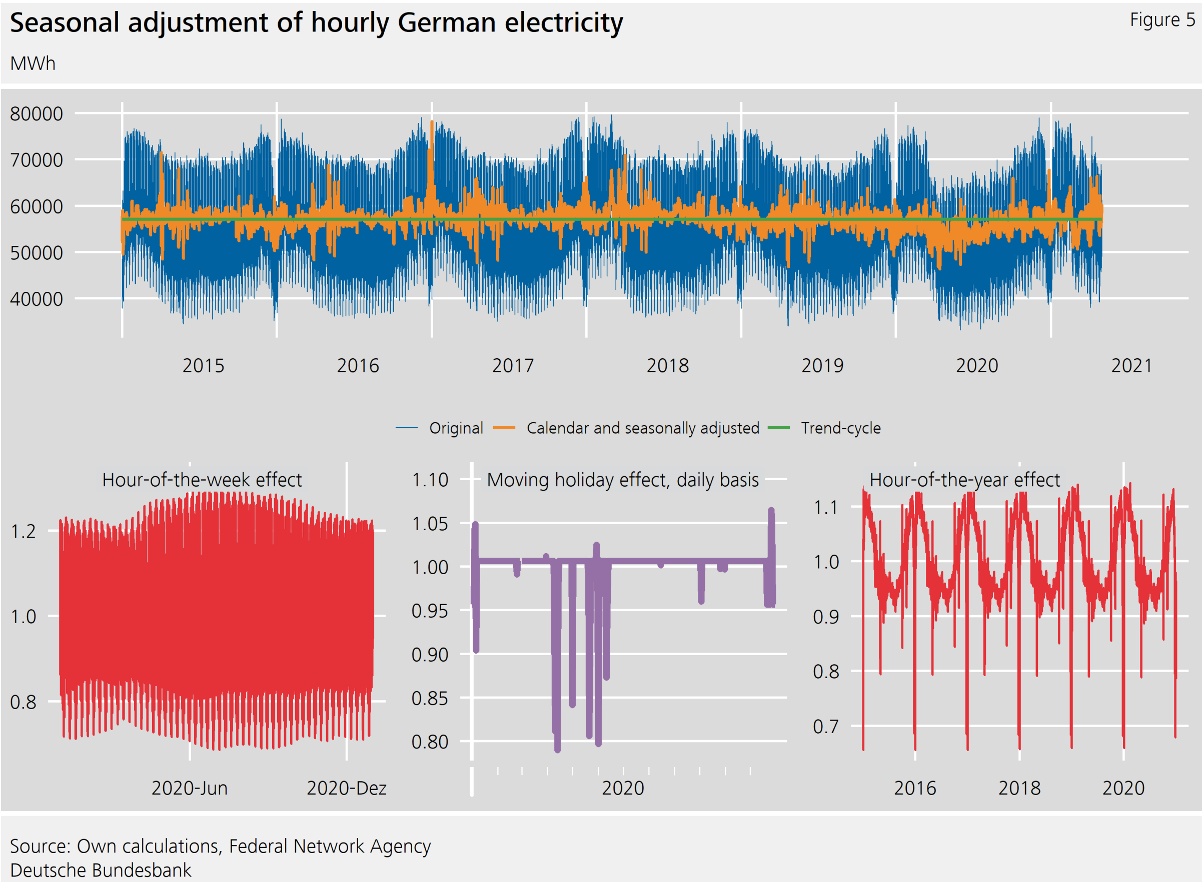

Figure 3.5: Seasonal adjustment of hourly German electricity

3.4.3 Google Trends: Unemployment

Weekly Google Trends data have to be downloaded in chunks of five years. Each of these chunks is a random sample of all search queries and therefore subject to noticeable revisions.[9] The chunks are chain-linked together and a value of 100 is added to avoid values too close to 0, which can lead to extreme seasonal and calendar factors if a multiplicative model is used.

[9] The average mean revisions are 7.14 points from week to week in the original series and 6.90 points in the calendar and seasonally adjusted data. The combined seasonal and calendar factor is on average revised by 0.74 percent in absolute terms.

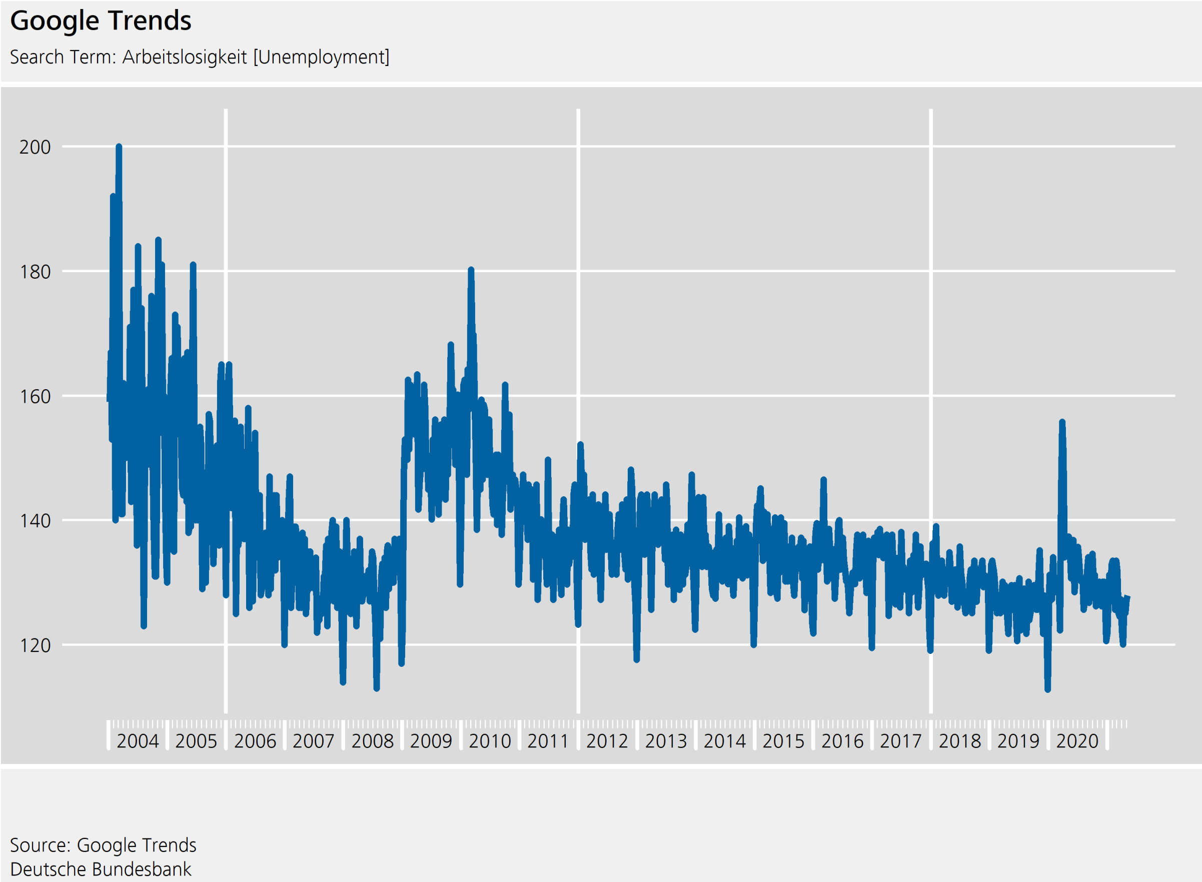

Figure 3.6: Google Trends

The series analysed here starts from 10 January 2004 and ends on 10 September 2021. The chunks of the series were downloaded on 13 September 2021.

The definition of the week does not adhere to the ISO 8601 standard. Instead it is defined to start on Sunday and end on Saturday. Such **date and time conventions** are of relevance for regressor construction. For example, Easter Sunday and Easter Monday fall in the same week, and their effect cannot be disentangled. In turn, we only need one regressor that captures the joint effect of Easter Sunday and Easter Monday.

For monthly or lower frequency time series, Christmas and New Year are seasonal effects as they fall in the same period each year. Depending on the modelling strategy, especially if the data are modified so that every year has 365 observations, this holds for daily data too, as these holidays always fall on the same day of the year. For weekly time series, as a consequence of their non-isochronicity and date conventions, Christmas' Day can fall in the 51st or 52nd week of the year, while New Year's Day falls in week 52, 53 or 1 (ISO 8601 standard) or week 0 or 1 (US convention).

The results of the RegARIMA model estimation are included in Table 3.4. As discussed, the construction of the data implies that some holidays always fall in the same week. Their impact is therefore indistinguishable. The inclusion of appropriate regressors for the Christmas and New Year's period can be challenging. Christmas Day always falls in the same week as either Christmas Eve or New Year's Eve. If both Christmas Eve and New Year's Eve are included as regressors, then, particularly for short series, separating the impact of Christmas Day from the other two days may be intricate, resulting in an unstable estimation.

Table 3.4: Estimated ARMA coefficients (in absolute terms) and moving holiday effects (in percent) for Google Trends, search term: Arbeitslosigkeit [unemployment]. A RegARIMA(0,1,1)(0,1,1) model is estimated.

	Estimate	S.E.		Estimate	S.E.
MA	−0.87	0.02	SMA	−0.75	0.04
Good Friday	−5.5	1.2	Christmas Eve	−6.7	1.7
Easter	−5.4	1.2	New Year's Eve	−5.3	1.6

Note: The week is defined as starting on Sunday and ending on Saturday. Thus, Easter covers both Sunday and Monday.

As can be seen from Figure 3.6 the series is very volatile, especially at the beginning of the series. It may be difficult to assess the seasonality of the time series visually, but seasonality tests[10] and the coefficients of the ARIMA model indicate seasonality. To obtain seasonal factors, we rely on the default settings of the fractional airline decomposition.

As the original data are revised considerably every week, the revision policy of this weekly series differs from the daily series discussed above. For the Google Trends series we follow a partial concurrent adjustment scheme, i.e. the calendar and seasonal components are re-estimated every week, but the order of the RegARIMA model is fixed and is only re-identified annually. The Appendix includes a discussion of graphical tools that can be used for weekly time series if a controlled current adjustment scheme is used.

3.5 Summary

The contribution of this paper is twofold. First, we performed an illustrative analysis of a small set of higher frequency time series. We discussed how these data differ from lower frequency time series and how this is relevant for seasonal adjustment in general and in light of the COVID-19 pandemic. Second, we developed a taxonomy of the central features of seasonal higher frequency time series. This list of features can contribute to the assessment of the seasonal adjustment of higher frequency time series and might serve as a building block in the development of quality diagnostics.

[10] We use the seasonality tests recommended by Ollech and Webel (2020) for monthly time series which are implemented in the {seastests} package in R. Setting the frequency of the time series to 52 ignores the non-integer type seasonality of the series. However, in practice this works well as an approximation. Both the QS and the Friedman test reject the null hypothesis of no seasonality at the 0.1-% level of significance. This also holds true if we only include the last five years of observations.

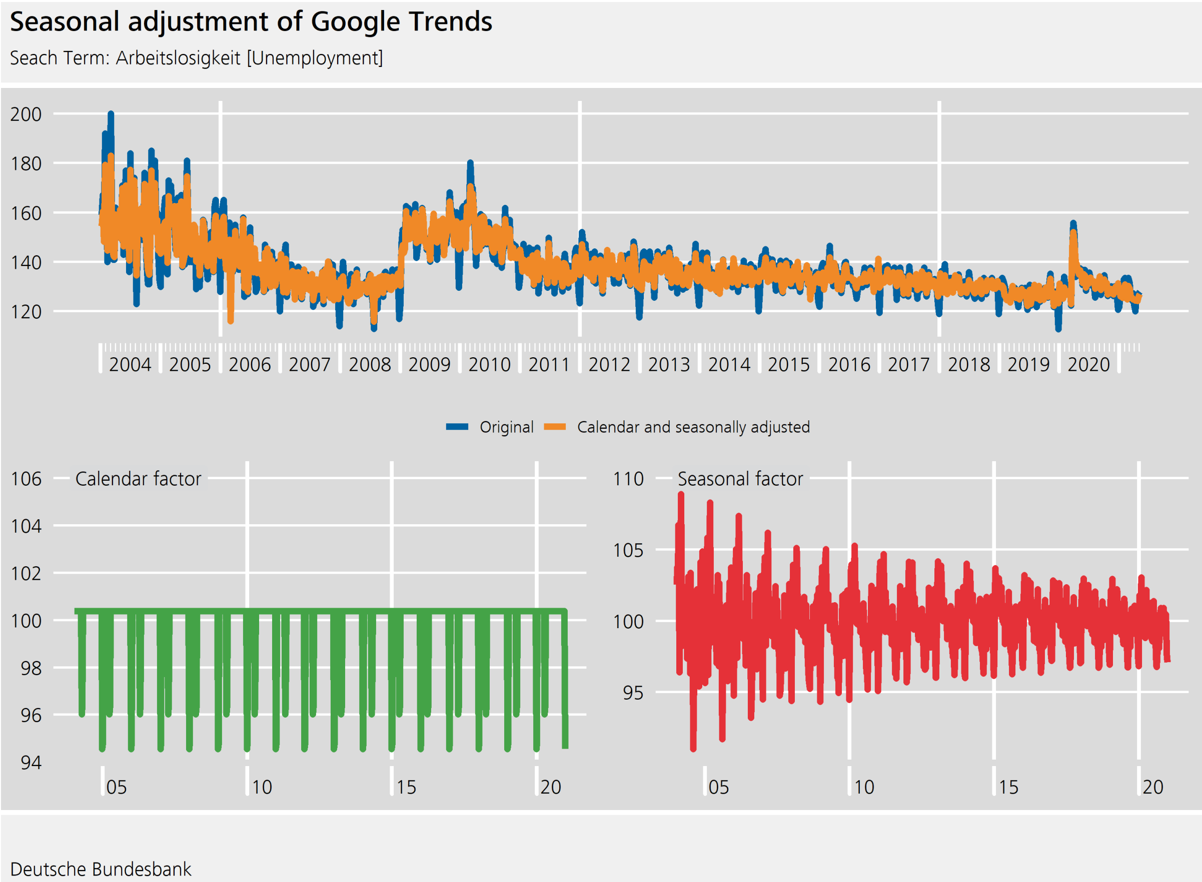

Figure 3.7: Seasonal adjustment of Google Trends

Further research may evaluate different procedures that allow the seasonal adjustment of higher frequency time series with respect to their ability to handle all of these features.

3.A Appendix

For a controlled current adjustment revision scheme, the results from the calendar and outlier adjustment are investigated and, perhaps most importantly, the previously forecasted seasonal components ($\hat{S}^F$) are compared to the new seasonal factors ($\hat{S}^N$) obtained after re-estimation. As reference points for the comparison, the raw seasonal factors ($\hat{S}^{raw}$), i.e. the combined seasonal-irregular component, are included in the analysis. The comparison is either based on tables or graphics, with the latter often being referred to as the SI ratios. The idea is to gauge whether $\hat{S}^N$ better captures the systematic development of $\hat{S}^{raw}$ compared to $\hat{S}^F$. If this is the case to a relevant extent, the previous components are discarded and the re-estimated results are used. Otherwise $\hat{S}^F$ can still be used and the previously published seasonally adjusted values are not revised.

For integer-type periodic influences the construction of the SI ratios is straightforward. Each subplot shows the $\hat{S}^N$, $\hat{S}^{raw}$ and $\hat{S}^F$ for a given period, e.g. a subplot for each month of the year. For series with a non-integer number of observations per year the

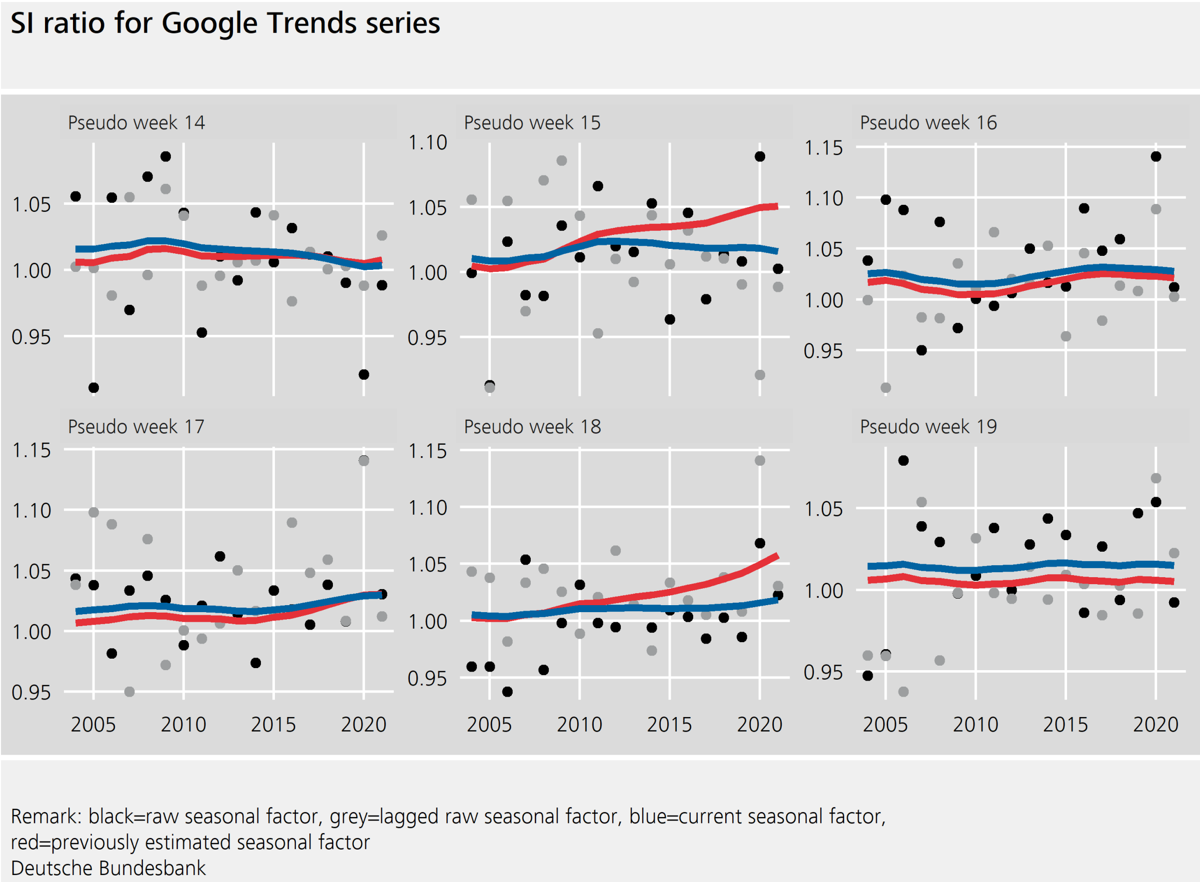

Figure 3.8: SI ratio for Google Trends series

configuration is more intricate. For weekly series, the estimated seasonal component of a given observation depends both on observations (multiples of) 52 weeks ago as well as 53 weeks ago and likewise on future observations. Therefore, $\hat{S}^{raw}$ should include all weeks that contribute to the estimation of the seasonal component. For the most recent observations, this is approximately the sequence of weeks with a step size of 52 and the respective previous week. This is shown by way of example in Figure 3.8. For any period of the seasonal component, the exact representation of the relevant raw seasonal factors is given by the sequence $(..., \hat{S}^{raw}_{t+\lceil 2\tau\rceil}, \hat{S}^{raw}_{t+\lfloor 2\tau\rfloor}, \hat{S}^{raw}_{t+\lceil\tau\rceil}, \hat{S}^{raw}_{t+\lfloor\tau\rfloor}, \hat{S}^{raw}_{t}, \hat{S}^{raw}_{t-\lfloor\tau\rfloor}, \hat{S}^{raw}_{t-\lceil\tau\rceil}, \hat{S}^{raw}_{t+\lfloor 2\tau\rfloor}, \hat{S}^{raw}_{t+\lceil 2\tau\rceil}, ...)$.

Bibliography

Arora, V. and Shi, S. (2016). Energy Consumption and Economic Growth in the United States. *Applied Economics*, 48(39):3763–3773.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6(1):3–73.

Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Cleveland, W. S. and Terpenning, I. J. (1982). Graphical methods for seasonal adjustment. *Journal of the American Statistical Association*, 77(377):52–62.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527.

Delle Monache, D., Emiliozzi, S., and Nobili, A. (2020). Tracking Economic Growth during the Covid-19: a Weekly Indicator for Italy. *Bank of Italy, Mimeo*.

Deutsche Bundesbank (2020). Seasonal Adjustment of the Daily Truck Toll Mileage Index. Methodology Report by the Seasonal Adjustment Section.

Do, L. P. C., Lin, K.-H., and Molnár, P. (2016). Electricity Consumption Modelling: A Case of Germany. *Economic Modelling*, 55:92–101.

Eraslan, S. and Götz, T. (2020). Weekly Activity Index for the German Economy. Available under URL https://www.bundesbank.de/wai.

Eurostat (2020). Guidance on Treatment of Covid-19-crisis Effects on Data.

Fenz, G. and Stix, H. (2021). Monitoring the Economy in Real Time with the Weekly OeNB GDP Indicator: Background, Experience and Outlook. *Monetary Policy & the Economy*.

Ladiray, D., Palate, J., Mazzi, G. L., and Proietti, T. (2018). Seasonal Adjustment of Daily and Weekly Data. In Mazzi, G. L., Ladiray, D., and Rieser, D. A., editors, *Handbook on Seasonal Adjustment*, chapter 29, pages 757–783. Publications Office of the European Union, Luxembourg.

Lewis, D. J., Mertens, K., and Stock, J. H. (2020a). Weekly Economic Index. URL https://www.newyorkfed.org/research/policy/weekly-economic-index.

Lewis, D. J., Mertens, K., Stock, J. H., and Trivedi, M. (2020b). Measuring Real Activity Using a Weekly Economic Index. *Federal Reserve Bank of New York*, 920.

Lewis, D. J., Mertens, K., Stock, J. H., and Trivedi, M. (2021). High-Frequency Data and a Weekly Economic Index During the Pandemic. In *AEA Papers and Proceedings*, volume 111, pages 326–330.

Lourenço, N. and Rua, A. (2021). The Daily Economic Indicator: Tracking Economic Activity Daily During the Lockdown. *Economic Modelling*, 100:1–10.

López Prol, J. and O, S. (2020). Impact of COVID-19 Measures on Short-Term Electricity Consumption in the Most Affected EU Countries and USA States. *iScience*, 23(10):1–29.

Ollech, D. (2018). Seasonal Adjustment of Daily Time Series. *Deutsche Bundesbank, Discussion Paper Series 41/2018*.

Ollech, D. (2021). Seasonal adjustment of daily time series. *Journal of Time Series Econometrics*, 13(2).

Ollech, D., Gonschorreck, N., and Hengen, L. (2021). Flexibilisation of X-11 for Higher-Frequency Data. In *NTTS Conference 2021*.

Ollech, D. and Webel, K. (2020). A random forest-based approach to identifying the most informative seasonality tests. Discussion Paper No 55/2020, Deutsche Bundesbank.

Proietti, T. and Pedregal, D. J. (2022). Seasonality in high frequency time series. *Econometrics and Statistics*.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Svetunkov, I. (2017). Statistical Models Underlying Functions of 'smooth' Package for R. Working paper, Lancaster University Management School.

U.S. Census Bureau (2016). *X-13ARIMA-SEATS Reference Manual Version 1.1.* Time Series Research Staff, Statistical Research Division, U.S. Census Bureau.

Webel, K. (2020). Challenges and Recent Developments in the Seasonal Adjustment of Daily Time Series. JSM Proceedings, Business and Economic Statistics Section, pages 1971–1997, Alexandria, VA: American Statistical Association.

Wegmüller, P., Glocker, C., and Guggia, V. (2021). Weekly Economic Activity: Measurement and Informational Content. Grundlagen für die Wirtschaftspolitik 17, Staatssekretariat für Wirtschaft SECO.

Woloszko, N. (2020). Tracking Activity in Real Time with Google Trends. OECD Economics Department Working Papers 1634.

CHAPTER 4

A random forest-based approach to identifying the most informative seasonality tests

Daniel Ollech and Karsten Webel

Abstract

Virtually every seasonal adjustment software includes an ensemble of tests for assessing whether a given time series is in fact seasonal and hence a candidate for seasonal adjustment. However, such tests are certain to produce either agreeing or conflicting results, raising the questions how to identify the most accurate tests and how to aggregate the results in the latter case. We suggest a novel random forest-based approach to answer these questions. We simulate seasonal and non-seasonal ARIMA processes that are representative of the macroeconomic time series analysed regularly by the Bundesbank. Treating the time series' seasonal status as a classification problem, we use the p-values of the seasonality tests implemented in the seasonal adjustment software JDemetra+ as predictors to train conditional random forests on the simulated data. We show that this aggregation approach avoids the size distortions of the JDemetra+ tests without sacrificing too much power compared to the most powerful test. We also find that the modified QS and Friedman tests are the most accurate ones in the considered ensemble.

Published as: Ollech, D. and Webel, K. (2023). A Random Forest-based Approach to Combining and Ranking Seasonality Tests. Journal of Econometric Methods, 12(1):117-130.

Author contributions: D.O. designed and implemented the seasonality tests and the random feature elimination algorithm. D.O analysed the results, K.W. assisted with the interpretation. D.O and K.W. wrote the manuscript.

Keywords: Binary classification; conditional inference trees; correlated predictors; simulation study; supervised machine learning

4.1 Motivation

To judge current economic developments and to forecast important target variables, such as quarterly gross domestic product, economists usually monitor a broad and well-established set of trustworthy key time series on a regular basis. Typically, some time series will enter this set in unadjusted form while other series will enter it in seasonally adjusted form, depending on the seasonal status of the particular time series. Examples include the data sets used by the Bundesbank for short-term forecasting of economic activity in Germany and other countries (Goetz and Hauzenberger, 2018; Pinkwart, 2018). Seasonal adjustment is thus an integral part of many economic analyses and business processes, including in particular case-by-case decisions on whether or not observed time series are in fact seasonal and should or should not be seasonally adjusted.

A variety of statistical tests has been developed over time to answer the latter question (e.g. Busetti and Harvey, 2003; Franses, 1992; Ghysels and Osborn, 2001). Each of those targets a specific manifestation of seasonality and for that reason many seasonal adjustment programmes implement more than just one test. For example, JDemetra+ (JD+) implements six seasonality tests[1] and Table 4.1 reports their outcomes for four monthly German macroeconomic time series that differ substantially in their seasonal behaviour, amongst other things (Figure 4.1). Retail trade turnover for games and toys displays a strong seasonal pattern that primarily originates from Christmas-related spikes in December figures. The harmonised index of consumer prices (HICP) for tobacco clearly lacks such strong seasonal movements as the series is dominated by a slowly rising trend that is interrupted by aperiodic level shifts associated with VAT increases. These visual impressions are unanimously confirmed by the JD+ tests. In contrast, neither the consumer price index (CPI) for energy nor the number of persons employed in the manufacturing of wearing apparel reveal their seasonal status as immediately as the other two series. Either series appears to be driven mainly by trend-cyclical and irregular movements, leaving any seasonal behaviour far less distinctive compared to the turnover series. Yet, some recurring intra-year movements, such as the minor troughs in the employment series in the middle of the years 2012 to 2014 and 2016, can still be eye-balled. The JD+ tests disagree: for the CPI series, four tests reject the null hypothesis of absence of seasonality at the 1% level of significance, while two tests fails to do so, albeit barely in one case; for the employment series, two tests provide strong evidence for presence of seasonality at the 1% level of significance, while one test provides mild evidence at the 5% level of significance, and the remaining tests do not reject the null hypothesis at this level.

This example illustrates two issues with any set of tests for the same testing problem. First, for some data the tests reach the same decision and we may ask whether the set

[1] Technical details about the tests are given in the Appendix.

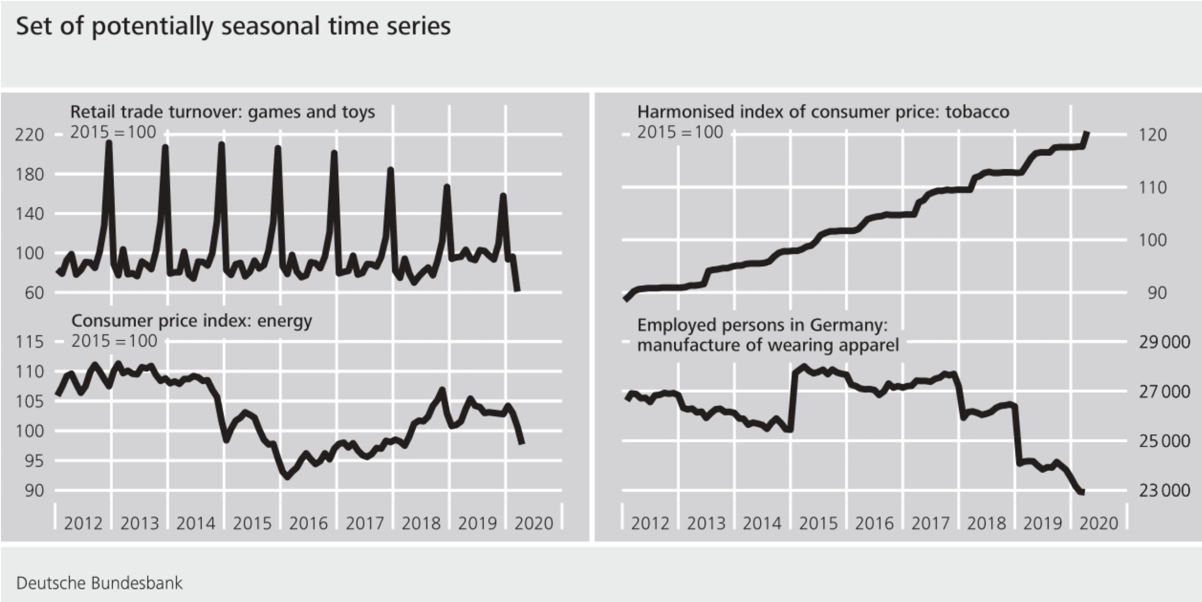

Figure 4.1: Four monthly macroeconomic time series for Germany.

is too large and if it can be reduced by eliminating seemingly redundant tests. Second, and more importantly, for some other data the tests reach different decisions and we may ask how their outcomes can be aggregated. Relatively simple aggregation rules include majority voting as well as Fisher-type χ^2 and union-of-rejection tests, the latter two of which have been considered especially for unit root (Harvey et al., 2009) and cointegration problems (Bayer and Hanck, 2013).

We take a different and—to the best of our knowledge—novel approach as we treat testing the seasonal status of a given time series as a classification problem with two classes and seek a random forest-based aggregation rule. To train random forests on representative data, we simulate non-seasonal and seasonal ARIMA models that closely resemble the macroeconomic time series stored in the Bundesbank's database. This is achieved through a specific simulation algorithm that is described as part of the data generation in Section 4.2. We then calculate the p-values of the JD+ seasonality tests from the simulated ARIMA models and use them as predictors for training random forests, together with a binary response that indicates the models' true seasonal status. This approach also facilitates a ranking of the seasonality tests with respect to the accuracy of their outcomes. Since the p-values display significant positive cross-correlations, rankings derived from importance metrics based on random forests with classical decision trees may be severely biased (Hothorn et al., 2006) and we thus opt for conditional inference trees. Section 4.3 briefly highlights the key differences between the two types of random forests. We next compare the size and power of the JD+ seasonality tests with the random forest classifications and show that the latter avoid the severe size distortions of the former without significant losses in power and that the modified QS and Friedman tests top the aforementioned accuracy ranking (Section 4.4). Our random forest-based approach is then applied to the four time series shown in Figure 4.1, disentangling especially the conflicting test outcomes for the CPI and employment series (Section 4.5). Finally, Section 4.6 concludes. All subsequent computations have been conducted in R (R Core Team, 2019).

4.2 Data Generation

We aim at simulating representative seasonal and non-seasonal time series that portray as accurately as possible the macroeconomic monthly data analysed regularly by the Bundesbank. To this end, ARIMA models will be our workhorse time series model since they cover a broad range of correlation structures, perform very well in forecasting exercises, and, as a consequence, have become the standard model in many seasonal adjustment approaches, especially those available in JD+.[2] An ARIMA model for a time series $\{y_t\}$ with τ observations per year is given by

$$\phi_p(B)\,\Phi_P(B^\tau)\,\nabla_1^d\,\nabla_\tau^D y_t = \theta_q(B)\,\Theta_Q(B^\tau)\,\varepsilon_t, \tag{4.1}$$

where B is the backshift operator, $B^k y_t = y_{t-k}$, $\nabla_k = 1 - B^k$, d and D indicate the non-seasonal and seasonal orders of differencing, ϕ_p and Φ_P are the non-seasonal and seasonal autoregressive (AR) polynomials of orders p and P, θ_q and Θ_Q are the non-seasonal and seasonal moving average (MA) polynomials of orders q and Q, and $\{\varepsilon_t\}$ is white noise with zero mean and finite variance. We will use the short-form notation $(pdq)(PDQ)$ to refer to a specific ARIMA model of type (4.1) and call it seasonal if $(PDQ) \neq (000)$ and non-seasonal if $(PDQ) = (000)$.

4.2.1 Model Selection

Representative ARIMA models are selected through an automatic modelling of monthly macroeconomic time series using the TRAMO (Gómez and Maravall, 2001) and regARIMA (Findley et al., 1998) approaches.[3] This includes an automatic identification of appropriate model orders $(pdq)(PDQ)$, estimation of ARMA parameters and detection of outliers for each single series. A total of 2,907 series which have been seasonally adjusted on a regular basis for many years is used to select the seasonal models. This data set is referred to as set SA (seasonally adjusted). It contains data on industrial production, orders received by industry, turnover in industry and retail trade, labour market statistics and consumer prices, with lengths ranging from 6 to 67 years. A total of 10,635 series which are not seasonally adjusted is used to select the non-seasonal models. These series

[2] JD+ implements the X-11 and ARIMA model-based approaches and their respective regARIMA and TRAMO pretreatment models, which are essentially linear time series regressions with ARIMA errors that may account for log transformations, outliers and calendar variation.

[3] The time series have been downloaded from the Bundesbank's database in January 2017.

are referred to as set NA (not adjusted). They contain data on national accounts, trade and capital markets, and have lengths between 4 and 72 years.

Both TRAMO and regARIMA identified a diverse set of ARIMA models, including pure AR, pure MA and mixed models. Let $\mathcal{M}(k)$ denote the set of identified models in category $k \in \{\mathrm{NA}, \mathrm{SA}\}$ and $p_{mk}^{(\mathrm{T})}$ and $p_{mk}^{(\mathrm{R})}$ the respective set-k shares of model $m \in \mathcal{M}(k)$. Each model's overall set-k share is then given by $p_{mk} = \left(p_{mk}^{(\mathrm{T})} + p_{mk}^{(\mathrm{R})}\right) \big/ 2$ and the model's simulation weight (rounded to 1 decimal place) is

$$w_{mk} = \frac{\tilde{p}_{mk}}{\sum_{j \in \mathcal{M}(k)} \tilde{p}_{jk}}, \tag{4.2}$$

where $\tilde{p}_{mk} = p_{mk} \cdot \mathbb{1}_{\left\{p_{mk}^{(\mathrm{T})} \geq 0.01 \,\mathrm{OR}\, p_{mk}^{(\mathrm{R})} \geq 0.01\right\}}$ is used to exclude rare models that have been identified for less than one percent of all time series.[4]

4.2.2 Model Simulation

We run Algorithm 1 for each selected representative ARIMA model $m \in \mathcal{M}(k)$ with $k \in \{\mathrm{NA}, \mathrm{SA}\}$ and each length $T \in \{60, 120, 240\}$. This algorithm combines the "NORmal-To-Anything" (NORTA) algorithm (Cario and Nelson, 1997) with logspline density estimation according to a particular knot addition and deletion algorithm (Stone et al., 1997). It simulates versions of estimated ARIMA models such that the simulated ARMA parameters closely resemble the multivariate distribution of the estimated ARMA parameters. This is achieved by imposing the restrictions that the simulated parameters

(R1) do not induce (additional) unit roots in the model's characteristic polynomial,

(R2) display the same correlation structure as the estimated parameters,

(R3) have univariate distributions with the same shape as the univariate distributions of the corresponding estimated parameters.

The basic idea is to initially draw sets of ARMA parameters from a multivariate Gaussian distribution such that the correlations of the simulated and estimated parameters are similar. The multivariate Gaussian density is then transformed into a set of univariate logspline densities. After elimination of ARMA parameters that induce additional unit roots, the correlations of the remaining parameters are compared to the correlations of the estimated ARMA parameters. In case of unacceptably high differences—which often result because transforming a Gaussian into a logspline density tends to reduce

[4]The shares and simulation weights of all models are reported in the supplementary material (Table 1), showing that some models $m \in \mathcal{M}(\mathrm{SA})$ were non-seasonal and we used $w_{m,\mathrm{SA}} = 0$ in such cases. The opposite "mis-identification" did not occur.

Algorithm 1 NORTA algorithm with logspline density estimation

Let n be the number of observed time series that all follow model (4.1) with the same orders $(pdq)(PDQ)$ and $\tilde{\nu}$ be the number of ARIMA models to be simulated under restrictions (R1)–(R3).

1: Set $m = p + q + P + Q$ and let $\hat{\mathbf{P}} \in \mathbb{R}^{m \times n}$ be the matrix of the estimated ARMA parameters. Calculate $\mathbf{\Sigma}_{\hat{\mathbf{P}}} \in \mathbb{R}^{m \times m}$, the correlation matrix of the parameters.
2: Apply logspline density estimation to each row of $\hat{\mathbf{P}}$ to obtain a non-parametric estimate $\hat{f}_j(\cdot)$ of the density of the j-th ARMA parameter, where $j \in \{1, \ldots, m\}$.
3: Set $\mathbf{\Sigma}_{\mathbf{P}}^{(1)} = \mathbf{\Sigma}_{\hat{\mathbf{P}}}$ to initialise the simulation of ARMA parameters, where $\mathbf{P} \in \mathbb{R}^{m \times \nu}$ denotes an empty matrix to be filled during the following loop.
4: **repeat**
5: In the i-th loop, simulate $\nu \gg \tilde{\nu}$ independent parameter vectors $\mathbf{P}_j \in \mathbb{R}^m$, where $\mathbf{P}_j \sim \mathcal{N}\left(\mathbf{0}_m, \mathbf{\Sigma}_{\mathbf{P}}^{(i)}\right)$ for each $j \in \{1, \ldots, n\}$. Set $\mathbf{P} = (\mathbf{P}_1 \ldots \mathbf{P}_\nu)$.
6: Set $\mathbf{Z} = (z_{jk}) \in \mathbb{R}^{m \times \nu}$, where $z_{jk} = \hat{F}_j^{-1}[\Phi(p_{jk})]$ for all $(j,k) \in \{1, \ldots, m\} \times \{1, \ldots, \nu\}$ and $\hat{F}_j(\cdot)$ and $\Phi(\cdot)$ are the distribution functions of $\hat{f}_j(\cdot)$ and the standard normal distribution, respectively.
7: Let $l \in \{0, \ldots, \nu\}$ be the number of columns of $\mathbf{Z}$ which contain ARMA parameters that induce additional unit roots. Remove the l columns from $\mathbf{Z}$ to obtain $\tilde{\mathbf{Z}} \in \mathbb{R}^{m \times (\nu - l)}$, the matrix of admissible ARMA parameters.
8: Select $\tilde{\nu}$ columns from $\tilde{\mathbf{Z}}$ according to simple random sampling without replacement, where $\tilde{\nu} \in \{1, \ldots, \nu - l\}$, this ensures that we get the correct number of parameters. Store the sampled columns in $\tilde{\mathbf{Z}}^{(\tilde{\nu})} \in \mathbb{R}^{m \times \tilde{\nu}}$.
9: Calculate $\mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}} \in \mathbb{R}^{m \times m}$, the correlation matrix of the sampled admissible ARMA parameters.
10: Calculate $\mathbf{\Delta} = |\mathbf{\Sigma}_{\hat{\mathbf{P}}} - \mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}}| = (\delta_{jk})$ and $\mathbf{C}_{\mathbf{\Delta}} = (c_{jk})$, where $c_{jk} = \mathbb{1}_{\{\delta_{jk} > \varepsilon\}}$ for all $(j,k) \in \{1, \ldots, m\}^2$ and some $\varepsilon > 0$.
11: **if** $\mathbf{C}_{\mathbf{\Delta}} \neq \mathbf{0}$ **then**
12: Set $\mathbf{\Sigma}_{\mathbf{P}}^{(i+1)} = \mathbf{\Sigma}_{\mathbf{P}}^{(i)} + \eta \left[\mathbf{\Sigma}_{\hat{\mathbf{P}}} - \mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}}\right] \odot \mathbf{C}_{\mathbf{\Delta}}$, where $\eta > 0$ and $\odot$ denotes the Hadamard product of two matrices, i.e. $\mathbf{A} \odot \mathbf{B} = (a_{jk} \cdot b_{jk})$.
13: **end if**
14: **until** $\mathbf{C}_{\mathbf{\Delta}} = \mathbf{0}$, or the maximum number of iterations is reached.
15: Simulate $\tilde{\nu}$ ARIMA models of order $(pdq)(PDQ)$ with the parameters stored in the columns of $\tilde{\mathbf{Z}}^{(\tilde{\nu})}$.

correlations between parameters—the entire algorithm is restarted with an appropriately modified correlation matrix of the initial multivariate Gaussian distribution.[5]

To exemplify the performance of Algorithm 1, Figure 4.2 shows kernel estimates of the univariate densities for the estimated and simulated parameters of the ARIMA model with the highest simulation weight: the (011)(011) model given by $\nabla_1 \nabla_\tau y_t = (1 -$

[5]In our application of Algorithm 1, we set $\nu = 100{,}000$, $\tilde{\nu} = \nu \cdot w_{mk}$, $\varepsilon = 0.02$ and $\eta = 0.5$, yielding a total of 600,000 simulated ARIMA models with Gaussian innovations. In Step 2, we restrict the density support to $[-1, 0.975]$ for AR parameters and to $[-0.99, 0.975]$ for MA parameters in seasonal and non-seasonal polynomials of order 1, and to $[-2, 2]$ for each parameter in non-seasonal AR and MA polynomials of orders 2 and 3. In Step 15, we let the length of the burn-in period for the simulation depend on the order of the finite AR representation of the ARMA polynomial, as in Briët et al. (2013).

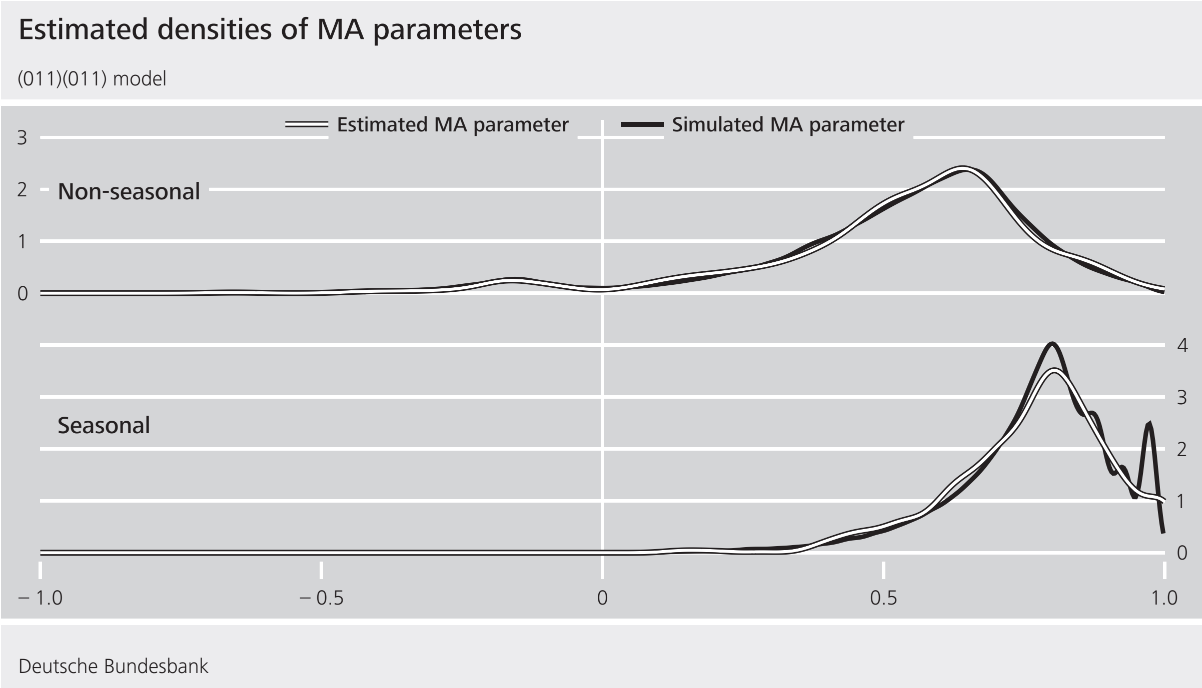

Figure 4.2: Gaussian kernel density estimates for the MA parameters in the (011)(011) model.

$\theta B)(1 - \Theta B^\tau)\, \varepsilon_t$. The estimated densities for the non-seasonal parameter θ are almost indistinguishable. The ones for the seasonal parameter Θ are also shaped very similarly in general.

4.2.3 Training and Validation Data

Let $\mathbf{X} = (\mathbf{X}_1 \dots \mathbf{X}_p)$ be a set of p predictors with $\mathbf{X}_j = (x_{1j} \dots x_{Nj})^\top$ for all $j \in \{1, \dots, p\}$ and $\mathbf{Y} = (y_1 \dots y_N)^\top$ be a vector of categorical responses with $y_i \in \{1, \dots, K\}$ for all $i \in \{1, \dots, N\}$. In our application, $\mathbf{X}$ will be given by the (asymptotic) p-values of five JD+ seasonality tests calculated from the 600,000 simulated ARIMA models[6] and $\mathbf{Y}$ will be a binary indicator for the true seasonal status of the simulated model. Hence, we will have $p = 5$, $N =$ 600,000 and $K = 2$. The training data is denoted by $\mathcal{L} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i = (x_{i1} \dots x_{ip})$ is the i-th (row) observation of the predictors $\mathbf{X}$. To evaluate the performance of the random forest classifiers, we also conveniently create via random sampling without replacement

- 50 independent "child" training data sets $\mathcal{L}^{(i)}$ from the "parent" $\mathcal{L}$ and
- 50 independent sampled validation data sets $\mathcal{V}_s^{(i)}$ from $\mathcal{V}^{(i)} = \mathcal{L} \setminus \mathcal{L}^{(i)}$, i.e. one $\mathcal{V}_s^{(i)}$ is sampled from each $\mathcal{V}^{(i)}$.

[6]No distribution theory has been proposed for the seasonal peaks (SP) test so far, thus no p-values are available. By unchangeable default in JD+, the simulated series are differenced once before the test calculation, trying to ensure stationarity.

The sample sizes are 7,500 for each $\mathcal{L}^{(i)}$ and 10,000 for each $\mathcal{V}_s^{(i)}$, which is mainly dictated by computational restrictions. We observed that computing more than 50 random forests did not alter the results but substantially increase the computational cost.

4.3 Random Forests

Our classification exercise will be based on random forests (RF) for three reasons: first, they provide highly accurate predictions in a broad range of real-world applications;[7] second, they have a non-"black box" tree-like structure that makes internal decision-making processes and the final classification comprehensible; third, they enable a ranking of the predictors in $\mathbf{X}$ with respect to their contribution to the final classification. In our application, however, the predictors, that is the p-values of five JD+ seasonality tests, exhibit high positive correlations that exceed 0.84 for any pair of predictors (Table 4.2), which is likely to induce a serious upwards bias in the estimated importance metrics for the predictors. Therefore, we consider RF with conditional inference trees that avoid such biases.[8]

Section 4.3.1 briefly summarises the basic concepts of classical RF. A detailed description can be found in Hastie et al. (2009). Section 4.3.2 highlights the key changes made for conditional inference trees.

4.3.1 Classical Approach

The classical RF algorithm (Breiman, 2001) utilises bootstrap aggregation, also known as bagging (Breiman, 1996; Bühlmann and Yu, 2002). In essence, it (1) draws bootstrap samples $\mathcal{L}_b$ of size N with replacement from the training data $\mathcal{L}$, where $b \in \{1, \dots, B\}$, (2) grows an unpruned binary classification tree $\mathcal{T}_b$ on each of those samples, and (3) aggregates the classifications of the single trees to a final classification, usually by majority vote or some weighted decision rule. During tree growing, a fresh sample $\tilde{\mathbf{X}}$ of size $\tilde{p} < p$ is drawn from $\mathbf{X}$ for the binary split of each node and the predictor that minimises node impurity within $\tilde{\mathbf{X}}$ is finally used to create the split. This shields RF from dominance of strong predictors and is repeated until node impurity cannot be reduced any further, or a pre-specified minimum size of terminal nodes is reached.

Since $\mathcal{L}_b$ is drawn with replacement, the "out-of-bag" (OOB) samples $\mathcal{O}_b = \mathcal{L} \setminus \mathcal{L}_b$ are non-empty and can be used for judging the RF performance by means of misclassification rates and for measuring the contribution of each predictor $\mathbf{X}_j$ to the final RF classification.

[7]Examples include forecasting stock price movements (Patel et al., 2015), diagnosing diseases (Hsieh et al., 2011), and detecting fraudulent emails (Almomani et al., 2013).

[8]We also investigated other machine learning algorithms, such as stochastic and adaptive boosting, support vector machines and feed forward neural networks, amongst others. The results are summarised in the supplementary material (Table 2) and further details are available upon request.

The latter can be achieved by calculating the mean decrease in prediction accuracy (MDA) after randomly permuting the values of $\mathbf{X}_j$ in the OOB samples, which essentially mimics absence of this predictor. Let $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)$ and $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)})$ denote the predicted classes of y_i obtained from $\mathcal{T}_b$ before and after this random permutation. The MDA is then given by

$$\mathrm{MDA}(\mathbf{X}_j) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathcal{O}_b|}\sum_{i\in\mathcal{O}_b}\left\{\mathbb{1}_{\{y_i\}}\left[\hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)\right] - \mathbb{1}_{\{y_i\}}\left[\hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)})\right]\right\}, \tag{4.3}$$

where $\mathbb{1}_{\mathcal{A}}[x] = 1$ if $x \in \mathcal{A}$, $\mathbb{1}_{\mathcal{A}}[x] = 0$ if $x \notin \mathcal{A}$ and $\pi(\cdot)$ is the permutation scheme.

4.3.2 Conditional Inference Trees

Correlations between the predictors in $\mathbf{X}$ are likely to introduce serious biases in (4.3) (Strobl et al., 2008, 2007), which is a consequence of the linkage between predictor selection and node splitting. In the absence of a truly influential predictor $\tilde{\mathbf{X}}_*$ in a sample $\tilde{\mathbf{X}}$, a predictor that is highly correlated with $\tilde{\mathbf{X}}_*$ but not with $\mathbf{Y}$ may be used to create the split and, hence, spuriously appear to be effective due to the neglected correlations.

The conditional inference framework for decision trees (Hothorn et al., 2006) untangles predictor selection and node splitting, which enables an ex ante exclusion of predictors that are not strongly related to $\mathbf{Y}$. Let each node n be represented by a N-dimensional vector $\mathbf{w}_n = (w_{n,1} \dots w_{n,N})^\top$ of integer case weights, where $w_{n,i} > 0$ if node n contains $(\mathbf{x}_i, y_i)$ and $w_{n,i} = 0$ otherwise. Predictor selection then starts by testing the global null hypothesis of no association between $\mathbf{Y}$ and each $\tilde{\mathbf{X}}_j$ in $\tilde{\mathbf{X}}$:

$$H_0^{(n)} = \bigcap_{j=1}^{\tilde{p}} H_0^{(n,j)} \quad \text{with} \quad H_0^{(n,j)} : \mathcal{D}(\mathbf{Y}|\tilde{\mathbf{X}}_j, \mathbf{w}_n) = \mathcal{D}(\mathbf{Y}|\mathbf{w}_n), \tag{4.4}$$

where $\mathcal{D}(\mathbf{Z})$ denotes the distribution of $\mathbf{Z}$. This hypothesis is rejected if the smallest (adjusted) p-value for rejecting the local null hypotheses $H_0^{(n,j)}$ is smaller than a prespecified nominal level α. In this case, the predictor $\tilde{\mathbf{X}}_{j*}$ with the strongest association to $\mathbf{Y}$ is identifiable from the respective local null hypothesis and eventually used to split node n. If (4.4) cannot be rejected, no split is made.

Applying the concept of conditioning also yields a version of (4.3) that is unbiased in the presence of correlated predictors (Strobl et al., 2008). To this end, $\pi(\cdot)$ is applied to the values of $\mathbf{X}_j$ only within subgroups of observations of $\mathbf{X}_j^c = (\mathbf{X}_1 \dots \mathbf{X}_{j-1}\mathbf{X}_{j+1} \dots \mathbf{X}_p)$.

This results in the conditional permutation scheme $\pi(\cdot)|\mathbf{X}^c_\cdot$ and the MDA is calculated as

$$\mathrm{MDA}(\mathbf{X}_j) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathcal{O}_b|}\sum_{i\in\mathcal{O}_b}\left\{\mathbb{1}_{\{y_i\}}\left[\hat{y}_i(\mathcal{T}_b,\mathbf{X}_j)\right] - \mathbb{1}_{\{y_i\}}\left[\hat{y}_i(\mathcal{T}_b,\mathbf{X}_{\pi(j)|\mathbf{X}^c_j})\right]\right\}, \tag{4.5}$$

where for each tree $\mathcal{T}_b$ the permutation grid for $\mathbf{X}_j$ is defined by the cut-points of $\mathbf{X}^c_j$ in $\mathcal{T}_b$.

4.4 Simulation Results

We grow a RF with conditional inference trees (henceforth referred to as conditional RF) on each "child" training data set $\mathcal{L}^{(i)}$, $i \in \{1,\ldots,50\}$. Cross-validation suggests forests of $B = 100$ trees. We sample $\tilde{p} = \lfloor\sqrt{p}\rfloor = 2$ predictors, i.e. p-values of JD+ seasonality tests, for node splitting and run the association tests of (4.4) on a nominal 5% level of significance, in line with common suggestions (Díaz-Uriarte and de Andrés, 2006; Strobl et al., 2007). For illustrative purposes, we also grow a classical RF on each $\mathcal{L}^{(i)}$, using essentially the same parameters and a minimum size of terminal nodes equal to 1.

Section 4.4.1 evaluates the size and power of the JD+ seasonality tests and the two RF approaches. Section 4.4.2 ranks the tests in descending order of accuracy as given by (4.5).

4.4.1 Size and Power

Table 4.3 reports the rejection rates of the "no seasonality" hypothesis for the seasonality tests and the two RF classifiers, which have been calculated from the 600,000 simulated ARIMA time series for the former and from the OOB samples $\mathcal{O}^{(i)}_b$ and sampled validation (VAL) data sets $\mathcal{V}^{(i)}_s$ for the latter.

Each test has acceptable power but suffers from upward size distortions, especially for the longer series. The modified QS test has the highest power regardless of the series' length, whereas the Friedman (FD) test has the least size distortions overall, although the modified QS test performs slightly better for the 5-year non-seasonal series and the seasonal peaks (SP) test does so for the 10- and 20-year non-seasonal series. It should also be noted that the size distortions of the modified QS test can in part be explained by the fact that this test is probably most sensitive to under-differencing, recalling that each test is applied to the once-differenced data. The modified QS test, for example, misclassifies 64.2% of the non-seasonal (021)(000) models when $\alpha = 0.01$, which already accounts for 1.47 percentage points of its rejection rate. In contrast, the other tests misclassify between 0.1% and 6.3% of those models. As suggested by an anonymous referee, we also looked at the rejection rates at the more granular levels of model categories and parameter categories for selected models. The former analysis revealed that the size distortions of the tests significantly increase with the complexity of the model, regardless of the chosen level of significance.[9] Amongst other things, we also studied the tests' performance for various seasonal models with $P = 1$ and/or $Q = 1$ and simulated parameters that induce near (seasonal) unit roots. However, these detailed analyses did not reveal significant differences in the rejection rates across parameter categories in any of the considered cases (the results are available upon request).

The RF classifiers do not display any size distortions at all—in fact, they are even slightly undersized at the 1% level of significance—but have to compromise some power compared to the most powerful seasonality test. Both effects are marginally more pronounced for conditional RF. The standard deviations of the rejection rates barely differ between the two classifiers, except for the 10- and 20-year seasonal series in the sampled VAL data for which they are slightly higher for conditional RF. Overall, the two RF classifiers have good size and power properties for any considered length of series.

Remark 10. Theoretical considerations suggest certain groups of seasonality tests with respect to the specific manifestation of seasonality tackled by the tests (Appendix). These conceptual specificities can be recognised, at least to some degree, in the correlations between the tests' p-values (Table 4.2) and rejection rates (Table 4.3). One group consists only of the modified QS test that is based on the sample autocorrelations and, hence, can handle both gradual changes in seasonality and volatile seasonal patterns very well. As a consequence, this test rarely misclassifies seasonal series. Another group consists of the FD, Kruskal-Wallis (KW), periodogram (PD) and seasonal dummies (SD) tests that primarily tackle stable seasonality, e.g. by checking differences between period-specific means. These tests tend to have lower size distortions but also slightly less power compared to the modified QS test. The last group consists only of the SP test that tackles visually significant spectral peaks at any seasonal frequency without an underlying distribution theory.

4.4.2 Accuracy Ranking of Seasonality Tests

We calculate the MDA measures (4.3) and (4.5) of each tests' p-values from each classical and conditional RF grown on $\mathcal{L}^{(i)}$ and visualise them with the aid of boxplots (Figure 4.3). For conditional RF, the modified QS test has the highest median MDA with a narrow margin over the FD test. This positioning may not come as a surprise given the high power of the former and the relatively low size distortions of the latter (Section 4.4.1). The SD, KW and PD tests finish far behind at third, fourth and fifth place, displaying a noticeably lower dispersion of the MDA measures.

[9] An example in which model complexity is measured through the sum of the non-seasonal ARIMA orders in (4.1) is given in the supplementary material (Table 3). Further details are available upon request.

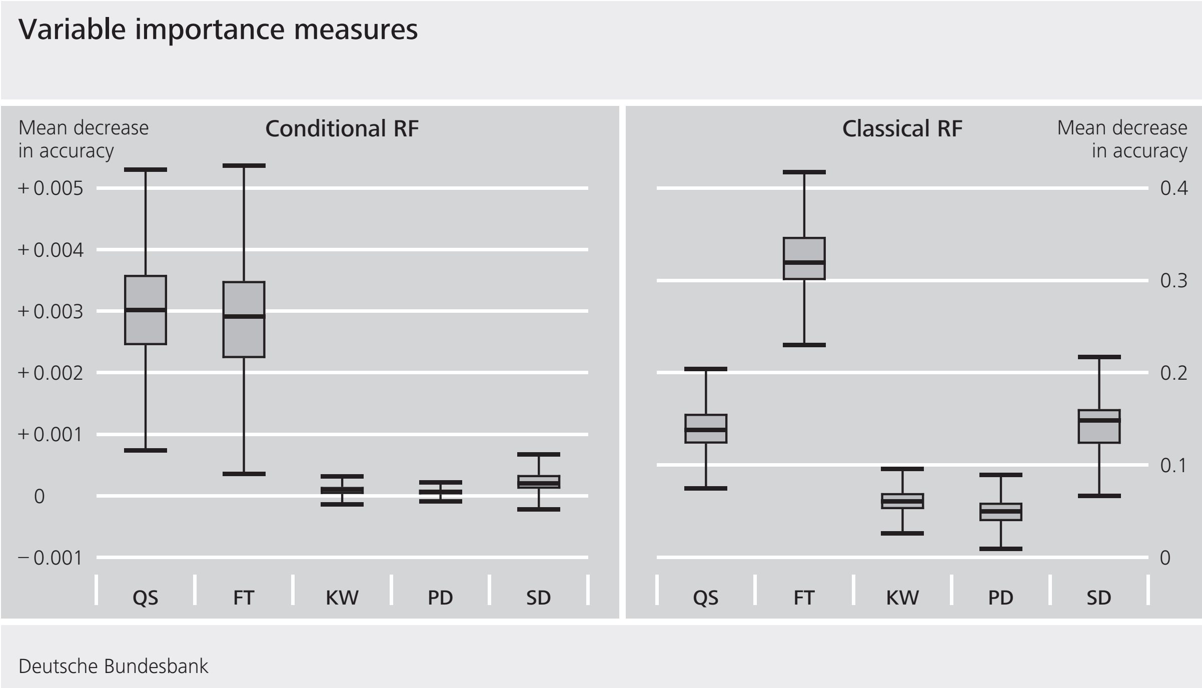

Figure 4.3: MDA measures (4.5) (*left*) and (4.3) (*right*) for five JD+ seasonality tests.

For classical RF, the FD test ranks first with a median MDA that is roughly twice as high as the measures of the second-ranked SD and third-ranked modified QS tests. Compared with conditional RF, this discrepancy in the accuracy rankings results from the biases introduced into (4.3) through the negligence of the p-values' correlations. In particular, the high positioning of the SD test raises suspicions given the test's large size distortions and moderate power relative to the other tests.

4.5 Application

We now return to the four German macroeconomic time series displayed in Figure 4.1 and demonstrate that conditional RF can also be used to disentangle the conflicting results of the JD+ seasonality tests. Recall that those unanimously classified retail trade turnover for games and toys as seasonal and HICP for tobacco as non-seasonal, but disagreed when judging CPI for energy and the number of persons employed in the manufacture of wearing apparel (Table 4.1).

Each of the 50 conditional RFs classifies CPI for energy as non-seasonal. Although contrasting with the majority vote of the tests, this decision seems reasonable from an economic perspective since more than 60% of the energy subcomponents, such as fuels and lubricants for personal transport equipment, heat energy and liquid fuels, are classified as non-seasonal by all JD+ seasonality tests. Each conditional RF also classifies the number of employed persons as seasonal, again in contrast to the majority vote of the JD+ tests (especially at the 1% level of significance), and finally yet rather unsurprisingly confirms the two concurrent classifications.

4.6 Summary

We treated the identification of the seasonal status of observed time series as a classification problem and showed that random forests (RF) with conditional inference trees can be used as a classifier. The training and validation data was composed of the p-values of five seasonality tests implemented in JDemetra+ as correlated predictors and a binary indicator of the seasonal status as the response. Both the p-values and binary indicators were calculated from a large set of simulated monthly ARIMA models that had been generated through a combination of the "NORmal-To-Anything" algorithm with logspline density estimation. This specific algorithm ensured the simulated data to be representative of the Bundesbank's macroeconomic time series database.

We reported severe size distortions and acceptable power for the seasonality tests and showed that the RF-based approach avoids size distortions without sacrificing too much power. Utilising the concept of mean decrease in accuracy, we also identified the modified QS and Friedman tests as the most accurate seasonality tests. An intuitive explanation may be that the Friedman test mainly targets stable seasonality while the modified QS test also allows for gradual change in seasonality over time and, hence, the two tests complement each other very well. Considering four macroeconomic time series for Germany, we finally illustrated the RF-based approach's capability of transforming conflicting results of individual seasonality tests into an unanimous classification.

Future research could extend the RF-based approach in a variety of ways. Making the training data more unbalanced with respect to seasonal status and length of the simulated data and/or adding quarterly, outlier-prone or borderline seasonal data could improve representativeness. Using alternative algorithms and methods, such as copulas, in the simulation process and/or faster RF-based classifiers, such as extremely randomised trees (Geurts et al., 2006), could improve computational efficiency. Last but not least, adding seasonality tests currently not implemented in JDemetra+ could further increase predictive power.

Appendix

We provide basic information about the six JD+ tests against the null hypothesis (H_0) of absence of seasonality in a weakly stationary time series $\{z_t\}$ of length T with τ observations per year. Non-stationary time series are made stationary by applying appropriate orders of (non-seasonal) differencing.

Modified QS Test

The modified QS test checks $\{z_t\}$ for significant positive autocorrelation at seasonal lags. Let $\gamma(h) = \mathbb{E}\left(z_{t+h}\, z_t\right) - \mathbb{E}^2\left(z_t\right)$ and $\rho(h) = \gamma(h)/\gamma(0)$ denote the lag-h autocovariance and autocorrelation, respectively, of $\{z_t\}$. Then, $H_0 : \rho(k) \leq 0$ for $k \in \{\tau, 2\tau\}$, and the QS-statistic is given by

$$QS = \left[T\left(T+2\right)\left(\frac{\hat{\rho}^2(\tau)}{T-\tau} + \frac{\left[\max\left\{0, \hat{\rho}(2\tau)\right\}\right]^2}{T-2\tau}\right)\right] \times \mathbb{1}_{(0,\infty)}\left[\hat{\rho}(\tau)\right],$$

where $\hat{\rho}(h)$ is the lag-h sample autocorrelation of $\{z_t\}$. The null distribution of QS is approximated by a χ^2-distribution with two degrees of freedom (Maravall, 2011).

Friedman Test

The Friedman (FD) test checks for significant differences between the period-specific mean ranks of the values of $\{z_t\}$. Let r_{ij} be the within-year rank of the observation in the i-th period of the j-th year, so that $1 \leq r_{ij} \leq \tau$, and $\mu_i = \mathbb{E}\left(r_{ij}\right)$. Then, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_\tau$, and the FD-statistic is calculated as a one-way ANOVA with repeated measures, which asymptotically follows a χ^2-distribution with $\tau - 1$ degrees of freedom under H_0 (Friedman, 1937).

Kruskal-Wallis Test

The Kruskal-Wallis (KW) test results from replacing the within-year ranks of the FD test with within-span ranks, so that $1 \leq r_{ij} \leq T$ and the KW-statistic is calculated as a one-way ANOVA without repeated measures. Hence, the same asymptotic null distribution applies (Kruskal and Wallis, 1952).

Periodogram Test

The periodogram (PD) test checks if a weighted sum of the spectral density $f(\omega) = (2\pi)^{-1} \sum_h \gamma(h)\, e^{-ih\omega}$ evaluated at the seasonal frequencies $\mathcal{S}(\tau) = \{\omega_1^\star, \ldots, \omega_{\tau/2}^\star\}$ with $\omega_j^\star = 2\pi j/\tau$ for $j \in \{1, \ldots, \tau/2\}$ is significantly different from zero. Let $\Sigma_{\mathcal{S}(\tau)} = 2\left[f(\omega_1^\star) + \cdots + f(\omega_{\tau/2-1}^\star)\right] + f(\omega_{\tau/2}^\star)$ be this weighted sum. Then, $H_0 : \Sigma_{\mathcal{S}(\tau)} = 0$, and the PD-statistic is obtained through plugging the periodogram estimates of $\{z_t\}$ into $\Sigma_{\mathcal{S}(\tau)}$ and proper rescaling, so that PD follows an F-distribution with $\tau - 1 - \mathbb{1}_{\mathcal{E}}[T]$ and $T - \tau + \mathbb{1}_{\mathcal{E}}[T]$ degrees of freedom under H_0, where $\mathcal{E}$ is the set of even integers.

Seasonal Peaks Test

The seasonal peaks (SP) test checks if $f(\omega)$ displays visually significant peaks within $\mathcal{S}(\tau)$. Let n_{vsp} be the number of such peaks and $SP = \mathbb{1}_{\{1,\dots,\tau/2\}}[n_{vsp}]$. Then, $H_0 : SP = 0$, and the SP-statistic is obtained from replacing n_{vsp} with a count estimator $\hat{n}_{vsp}$ that aggregates information from the Tukey and AR(30) estimators of $f(\omega)$. Criteria for visual significance, including simulated critical values, and the aggregation rule are derived in Maravall (2011).

Seasonal Dummies Test

Dropping the stationarity assumption on $\{z_t\}$, the seasonal dummies (SD) test checks if the joint effects of the $\tau - 1$ seasonal dummies, denoted by $\boldsymbol{\beta} \in \mathbb{R}^{\tau-1}$, are significantly different from zero in a time series regression with a constant mean and ARIMA $(pdq)(000)$ disturbances. Thus, $H_0 : \boldsymbol{\beta} = \mathbf{0}$, and the SD-statistic is the usual F-statistic based on the OLS estimator of $\boldsymbol{\beta}$, which follows an F-distribution with $\tau - 1$ and $T - d - p - q - \tau - 1$ degrees of freedom under H_0 (Gómez and Maravall, 2001).

Table 4.1: Test statistics and p-values of the JD+ seasonality tests for the time series shown in Figure 4.1.

	Retail trade turnover: games and toys		HICP: tobacco		CPI: energy		Employed persons: manufacture of wearing apparel	
Test	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
QS	537.787	0.000	0.750	0.687	7.463	0.024	6.551	0.038
FD	236.337	0.000	13.031	0.291	29.804	0.002	46.864	0.000
KW	258.577	0.000	5.363	0.912	33.755	0.000	44.127	0.000
SP	AT AT AT		⋆⋆ a⋆ ⋆⋆		⋆t ⋆⋆ ⋆⋆		⋆⋆ A⋆ ⋆⋆	
	AT AT AT		⋆⋆ ⋆⋆ a⋆		⋆⋆ a⋆ ⋆⋆		⋆⋆ ⋆⋆ ⋆⋆	
PD	323.551	0.000	0.376	0.965	3.140	0.001	1.796	0.062
SD	364.898	0.000	0.359	0.970	3.031	0.001	1.747	0.071

Remarks: All series have been differenced once, the CPI series has additionally been logged, as suggested by the automatic log-level-test (Gómez and Maravall, 2001). The tests are explained in the Appendix. For the SP test, a lower (upper) case letter refers to a visually significant peak at the 10% (1%) level of significance in the estimated Tukey (T) and AR(30) (A) spectra at the seasonal frequency that corresponds to the letters' position.

Table 4.2: Spearman correlations between the p-values of the JD+ seasonality tests.

	QS	FD	KW	PD	SD	SP
QS	1.000					
FD	0.875	1.000				
KW	0.849	0.942	1.000			
PD	0.844	0.941	0.972	1.000		
SD	0.844	0.921	0.946	0.944	1.000	
SP	0.870	0.839	0.825	0.814	0.828	1.000

Remarks: The tests have been applied to the 600,000 simulated ARIMA time series (pairwise complete cases). The color code indicates certain groups of seasonality tests (Remark 10). Other correlation measures, such as Bravais-Pearson coefficients, yield very similar results.

Table 4.3: Rejection rates of the "no seasonality" hypothesis (as a percentage) for the JD+ seasonality tests and the RF approaches.

Classifier			Non-seasonal series All	5-year	10-year	20-year	Seasonal series All	5-year	10-year	20-year
Conditional RF	OOB	Mean	0.76	0.63	0.83	0.82	97.92	97.62	98.01	98.13
		Std*	0.16	0.25	0.25	0.21	0.17	0.38	0.37	0.29
	VAL	Mean	0.77	0.66	0.82	0.82	97.87	97.65	97.95	98.02
		Std*	0.17	0.22	0.24	0.28	0.24	0.33	0.39	0.42
Classical RF	OOB	Mean	0.84	0.75	0.88	0.89	98.09	97.74	98.18	98.34
		Std*	0.16	0.27	0.22	0.25	0.16	0.35	0.34	0.29
	VAL	Mean	0.82	0.74	0.85	0.88	98.02	97.75	98.10	98.23
		Std*	0.18	0.24	0.24	0.30	0.19	0.32	0.34	0.36
Seasonality tests	$\alpha = 0.01$	QS	4.84	2.48	4.99	7.07	98.51	98.25	98.59	98.71
		FD	2.08	1.46	2.25	2.52	97.91	97.75	98.05	97.93
		KW	2.37	1.86	2.56	2.70	96.22	96.10	96.32	96.25
		PD	3.21	3.18	3.30	3.16	96.35	96.58	96.40	96.09
		SD	4.04	4.35	4.09	3.67	97.30	97.50	97.31	97.09
	$\alpha = 0.05$	QS	7.42	4.92	7.52	9.82	98.78	98.61	98.89	98.85
		FD	6.64	5.63	6.96	7.33	98.42	98.38	98.50	98.36
		KW	6.91	6.24	7.12	7.36	96.85	96.87	96.93	96.76
		PD	8.14	8.31	8.15	7.95	96.84	97.16	96.89	96.49
		SD	9.11	9.68	9.19	8.46	97.76	97.97	97.78	97.52
		SP	$*$	$*$	6.63	5.04	$*$	$*$	98.26	98.05

$\star$ = in percentage points. Std = standard deviation.

Remarks: The SD test is run with $(pdq) = (011)$. The SP cannot be calculated for monthly data with less than seven years of observations. It is thus run only for the 10- and 20-year series, using a nominal 10% level of significance for detecting visually significant peaks in the estimated Tukey and AR(30) spectra. Due to the absence of a distribution theory, its rejection rates are unaffected by the choice of α. The color code indicates certain groups of seasonality tests (Remark 10).

Bibliography

Almomani, A., Gupta, B., Atawneh, S., Meulenberg, A., and Almomani, E. (2013). A Survey of Phishing Email Filtering Techniques. *IEEE Communications Surveys & Tutorials*, 15(4):2070–2090.

Bayer, C. and Hanck, C. (2013). Combining non-cointegration tests. *Journal of Time Series Analysis*, 34(1):83–95.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Briët, O. J., Amerasinghe, P. H., and Vounatsou, P. (2013). Generalized Seasonal Autoregressive Integrated Moving Average Models for Count Data with Application to Malaria Time Series with Low Case Numbers. *PloS ONE*, 8(6):1–9.

Bühlmann, P. and Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics*, 30(4):927–961.

Busetti, F. and Harvey, A. C. (2003). Seasonality Tests. *Journal of Business & Economic Statistics*, 21(3):420–436.

Cario, M. C. and Nelson, B. L. (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Díaz-Uriarte, R. and de Andrés, S. A. (2006). Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics*, 7:Article 3.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., and Chen, B.-C. (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. *Journal of Business & Economic Statistics*, 16(2):127–152.

Franses, P. H. (1992). Testing for Seasonality. *Economics Letters*, 38(3):259–262.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Ghysels, E. and Osborn, D. R. (2001). *The Econometric Analysis of Seasonal Time Series*. Cambridge University Press, Cambridge.

Goetz, T. B. and Hauzenberger, K. (2018). Large mixed-frequency VARs with a parsimonious time-varying parameter structure. Discussion Paper No 40/2018, Deutsche Bundesbank, Frankfurt.

Gómez, V. and Maravall, A. (2001). Automatic Modeling Methods for Univariate Series. In Peña, D., Tiao, G. C., and Tsay, R. S., editors, *A Course in Time Series Analysis*, pages 171–201. Wiley, New York.

Harvey, D. I., Leybourne, S. J., and Taylor, A. R. (2009). Simple, robust, and powerful tests of the breaking trend hypothesis. *Econometric Theory*, 25(4):995–1029.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. Springer, Heidelberg, second edition.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hsieh, C.-H., Lu, R.-H., Lee, N.-H., Chiu, W.-T., Hsu, M.-H., and Li, Y.-C. J. (2011). Novel Solutions for an Old Disease: Diagnosis of Acute Appendicitis with Random Forest, Support Vector Machines, and Artificial Neural Networks. *Surgery*, 149(1):87–93.

Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Maravall, A. (2011). *Seasonality Tests and Automatic Model Identification in TRAMO-SEATS*. Bank of Spain. Mimeo.

Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015). Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques. *Expert Systems with Applications*, 42(1):259–268.

Pinkwart, N. (2018). Short-term forecasting economic activity in Germany: a supply and demand side system of bridge equations. Discussion Paper No 36/2018, Deutsche Bundesbank, Frankfurt.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial Splines and their Tensor Products in Extended Linear Modeling. *The Annals of Statistics*, 25(4):1371–1470.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9:Article 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8:Article 25.

CHAPTER 5

An overall seasonality test derived from recursive feature elimination in conditional random forests

Daniel Ollech and Karsten Webel

Abstract

Many seasonal adjustment programs contain a set of statistical tests to check observed time series for presence of seasonality. Since these tests will not always reach a consensus, some programmes also implement an overall seasonality test that condenses the outcomes of the single tests into an unambiguous decision. However, the design of such overall tests is usually not justified from a theoretical perspective but derived from empirical criteria and thus reflects to a large degree the experience and/or preferences of the designer. We propose a tree-like overall seasonality test based on a two-step procedure. In the first step, we identify the most informative tests from a large set of candidates by applying recursive feature elimination in conditional random forests. In a second step, we design a single conditional inference tree from the identified tests to achieve the classification accuracy of the random forests whilst at the same time providing interpretable results. We show that a sensible decision rule can be derived by combining the Kruskal-Wallis test with two variants of the modified QS test. The resulting overall seasonality test has low type I and II misclassification rates and avoids excessive complexity.

Published as: Ollech, D. and Webel, K. (2018). An Overall Seasonality Test Based on Recursive Feature Elimination in Conditional Random Forests. In Proceedings of the 5th International Conference on Time Series and Forecasting:20–31.[1]

[1] After submission of this thesis, a revised version has been published as Ollech, D. and Webel, K. (2025). A Novel Tree-Based Combined Test for Seasonality. Data Science in Science, 4(1),1–11.

Author contributions: D.O. designed and implemented the simulation study. D.O. designed and implemented the seasonality tests and the Random Forest based analysis. D.O analysed the results, K.W. assisted with the interpretation. D.O and K.W. wrote the manuscript.

Keywords: Binary classification; correlated predictors; JDemetra+; simulation study; supervised machine learning; test combination

5.1 Motivation

Seasonal behaviour can manifest itself in time-series data in a variety of ways, ranging from significant differences between the period-specific averages of the detrended data over large positive sample autocorrelations at seasonal lags to spectral peaks at seasonal frequencies. Accordingly, a vast range of seasonality tests has been developed over time and usually each of these tests targets a particular manifestation of seasonality. Practitioners thus often pay attention to an ensemble of tests rather than a single test, diversifying in some sense the risk of misjudgement. This strategy works well as long as the tests agree, which is usually the case for the majority of time series. In some cases, however, the tests yield conflicting results, which happens most likely for time series that are not textbook examples of seasonal or non-seasonal data. In such a case, overall seasonality tests can be quite useful. Such tests condense the outcomes of several seasonality tests into a final unambiguous decision on the seasonal status of the observed time series.

Many popular seasonal adjustment programmes contain an overall test. For example, the X-11 method implements a combined test for presence of identifiable seasonality (Morry and Lothian, 1978), which is also available in JDemetra+ (JD+). This test utilises a tree-like concept as it runs a sequence of seasonality tests in a pre-specified order at pre-specified levels of significance. Figure 5.1 shows the decision rule. Similarly, TRAMO/SEATS implements an overall test based upon a simple aggregation rule (Maravall, 2011). It considers the modified QS test, the non-parametric Friedman test, the test for seasonal peaks and the F-test on seasonal dummies (see the overview in Ollech and Webel, 2020)—each at the 1% level of significance—and deems the observed series (1) seasonal if at least two tests indicate presence of seasonality; (2) weakly seasonal if only the QS test or only the F-test on seasonal dummies indicates presence of seasonality; (3) non-seasonal in any other case. This rule performs reasonably well for simulated data, despite a slight bias towards over-detection of seasonality. It also affects the choice of initial parameters in automatic model identification.

However, to the best of our knowledge, the design of these (and other) overall tests has not been formulated from a theoretical perspective so far. In particular, set of considered tests, order of testing and chosen levels of significance have not been justified theoretically. Instead, it seems that those ingredients have been determined by educated guessing or simulations and thus reflect something that works well empirically. More sophisticated combinations are warranted. The previous chapter provides a proof of concept for the CRF approach to identifying informative seasonality test. The overall test derived in the present paper is based upon a recursive application of this concept in order to find a good combination.

The remainder of this paper is organised as follows. Section 5.2 presents conditional random forests and section 5.3 presents seasonality tests. Section 5.4 discusses the vari-

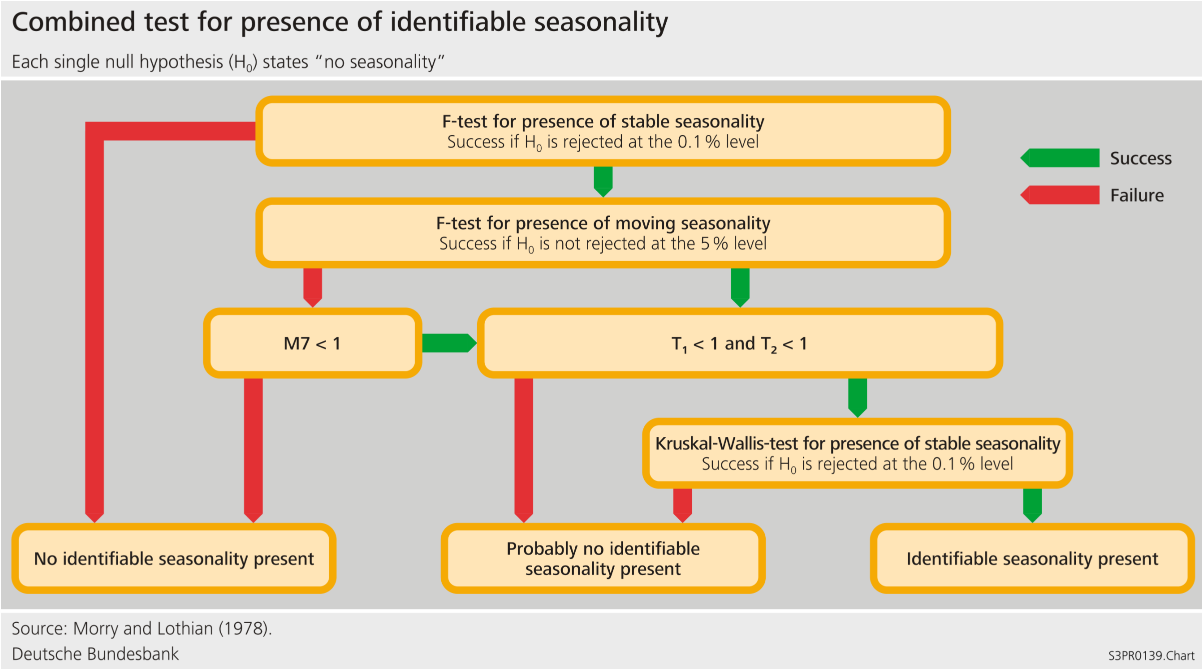

Figure 5.1: Overall seasonality test of the X-11 method. The heuristic measures are defined as $T_1 = \min\{7/F_S, 9\}$, $T_2 = \min\{3F_M/F_S, 9\}$ and $M7 = \sqrt{(T_1 + T_2)/2}$, where F_S and F_M are the statistics of the F-tests for presence of stable and moving seasonality.

able selection and construction of our overall test, section 5.5 presents the results from a simulation study and section 5.6 from a real world application. Section 5.7 discusses the overall test. Finally, section 5.8 concludes.

5.2 Conditional random forests

Random forests are essentially a collection of classification or regression trees that differ as they are built upon different bootstrap samples from the training data. To further induce randomisation and prevent any predictor from dominating the construction of every tree, only a random sample of the regressors are considered as splitting variable at each node (for an extensive discussion of random forests, see Ollech and Webel, 2020).

Conditional random forests (CRF) differs from random forests as developed by Breiman (2001) in two respects: variable selection and variable importance measures (VIMs). Let $\mathcal{L} = (\mathbf{X}\mathbf{Y})$ be the training data, where $\mathbf{X} = (\mathbf{X}_1 \dots \mathbf{X}_p)$ is a set of p predictors with $\mathbf{X}_j = (x_{1j}, \dots, x_{Nj})^\top$ for all $j \in \{1, \dots, p\}$ and $\mathbf{Y} = (y_1, \dots, y_N)^\top$ is a vector of categorical responses with $y_i \in \{1, \dots, K\}$ for all $i \in \{1, \dots, N\}$. For $b \in \{1, \dots, B\}$, a bootstrap sample $\mathcal{L}_b$ is drawn with replacement from $\mathcal{L}$, and an unpruned decision tree $\mathcal{T}_b$ is grown for the sample within the conditional inference framework (Hothorn et al., 2006). As opposed to the classical approach, the binary split of any terminal node is created according to the following generic algorithm:

1. Test the global null hypothesis of pairwise independence between $\mathbf{Y}$ and each predictor $\mathbf{X}_j$. Stop if this hypothesis cannot be rejected. Otherwise, identify the predictor $\mathbf{X}_{j*}$ with the strongest association to $\mathbf{Y}$ by means of standardised linear statistics within the permutation test framework of Strasser and Weber (1999).

2. Take $\mathbf{X}_{j*}$ as splitting variable. Find the optimal binary split of the node using pre-specified splitting criteria, such as two-sample linear statistics (Hothorn et al., 2006), that are in line with the criteria applied in the previous step.

Tree growing is stopped if a pre-specified minimum number of observations in terminal nodes, n_{min}, is reached or the stop criterion mentioned in Step 1 of the generic algorithm applies. Finally, the forest classification is given by the unweighted majority vote of the tree classifications.

As for the classical approach, subsets of $\mathcal{L}$ can be used for validation purposes. Let $\mathcal{O}_b = \mathcal{L} \setminus \mathcal{L}_b$ be the "out-of-bag" (OOB) data of the b-th bootstrap sample, that is the training data not selected in $\mathcal{L}_b$. The CRF's performance can then be judged by means of misclassification rates in the OOB samples. Alternatively, external validation (VAL) data can be considered, i.e. data that have not been considered at all for the construction of the forest.

The misclassification rates in the OOB samples can also be used to evaluate the importance of any predictor $\mathbf{X}_j$ in terms of its contribution to a given forest's classification. The basic idea is to permute the values of $\mathbf{X}_j$ in the OOB samples in order to mimic absence of $\mathbf{X}_j$. Then, the difference between the misclassification rates before and after permutation yields the the mean decrease in accuracy of $\mathbf{X}_j$ when absent. However, this VIM is likely to be biased if a purely random permutation scheme is applied in the presence of correlated predictors. A conditional permutation scheme (Strobl et al., 2007, 2008) prevents ex ante the overestimation of seemingly influential predictors $\mathbf{X}_j$ that in fact are not strongly associated with $\mathbf{Y}$ but appear as such due to a high correlation with a truly influential predictor. Unconditionally permuting $\mathbf{X}_j$ will eliminate the relationship with $\mathbf{Y}$ as well as with $\mathbf{X}_i$ for all $i \neq j$. The latter relationship prevails by conditionally permuting $\mathbf{X}_j$. To this end, the values of $\mathbf{X}_j$ are randomly permuted only within subgroups of its complement $\mathbf{X}_j^c = (\mathbf{X}_1 \ldots \mathbf{X}_{j-1}\mathbf{X}_{j+1} \ldots \mathbf{X}_p)$, resulting in the conditional permutation scheme $\pi(\cdot)|\mathbf{X}_\cdot^c$. Let $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)$ and $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)|\mathbf{X}_j^c})$ denote the predicted classes of y_i obtained from $\mathcal{T}_b$ before and after conditional permutation of the values of $\mathbf{X}_j$ in $\mathcal{O}_b$. The respective VIM is then given by

$$\mathrm{VI}\left(\mathbf{X}_j\right) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{O}_b}\left[\frac{\mathbb{1}_{\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)|\mathbf{X}_j^c})\}}}{|\mathcal{O}_b|} - \frac{\mathbb{1}_{\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)\}}}{|\mathcal{O}_b|}\right], \tag{5.1}$$

where for each tree $\mathcal{T}_b$ the permutation grid for $\mathbf{X}_j$ is defined by the cut-points of $\mathbf{X}_j^c$ in $\mathcal{T}_b$, i.e. by the binary splits determined by the classification or regression trees.

5.3 Seasonality tests

Let $\{z_t\}$ denote a weakly stationary time series of length T with τ observations per year. Whenever appropriate, the notation will be changed to $\{z_{ij}\}$, where z_{ij} is the observation of the i-th period, $i \in \{1, \dots, \tau\}$, in the j-th year. Furthermore, let $(pdq)(PDQ)$ abbreviate the ARIMA model

$$\phi_p(B)\, \Phi_P(B^\tau)\, \nabla_1^d\, \nabla_\tau^D x_t = \theta_q(B)\, \Theta_Q(B^\tau)\, \varepsilon_t,$$

where B is the Backshift operator, $B^k x_t = x_{t-k}$, $\nabla_k = 1 - B^k$, ϕ_p and Φ_P are the non-seasonal and seasonal autoregressive operators, θ_q and Θ_Q are the non-seasonal and seasonal moving average operators, and $\{\varepsilon_t\}$ is white noise with zero mean and finite variance. Then, a given ARIMA model is said to be non-seasonal (N-S) if $(PDQ) = (000)$ and seasonal (S) in any other case.

5.3.1 Candidate tests

We consider three branches of seasonality tests, each of which states absence of seasonality in $\{z_t\}$ as the null hypothesis:

- The six seasonality tests implemented in JD+ have been described in detail by Ollech and Webel (2020), along with respective intuitions. We consider five of them: the modified QS test (QS), the Friedman test (FT), the Kruskal-Wallis test (KW), the periodogram test (PD) and the F-test on seasonal dummies (SD). A seasonal pattern usually manifests itself in positive autocorrelation at seasonal lags. Therefore, the QS test checks whether this seasonal autocorrelation is significantly larger than zero. Both the FT and KW test are one-way ANOVAs with repeated resp. without repeated measures on the ranks of $\{z_t\}$. SD includes seasonal dummies in a RegARIMA model and tests whether their effects are simultaneously zero. The PD test estimates a weighted sum of the periodogram evaluated at the seasonal frequencies and tests if this sum is significantly different from zero.

- We also consider selected tests from the literature on medical statistics, which are primarily used for detection of periodic occurrences of diseases, and a seasonal unit root test. More detailed information on these tests is provided below.

- The following tests are also applied to the residuals obtained from fitting a non-seasonal ARIMA model to the unadjusted data, where (pdq) is determined by the automatic model identification routine Hyndman and Khandakar (2008) with $p \leq 3$, $q \leq 3$ and $p + q \leq 3$[2]: the QS, FT, KW, Roger, Welch and Welch-Kruskal-Wallis

[2] These restrictions are aimed at ensuring that the high-order of the non-seasonal ARIMA model does not extract seasonal movements from the time series. Also, avoid over-fitting.

tests. The short form notation of each of these tests is the short form notation of the respective original test plus the suffix "-R".

Welch test

Welch (1951) developed an ANOVA test without repeated measures for data with heterogeneous variance. Basically, it allows for heteroskedastic seasonality by scaling the observations of each period by the period-specific sample variance s_i and the number of observations n_i in period i. For the i-th period in year j, let $\mu_i = \mathbb{E}(z_{ij})$. Then, the null hypothesis is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_\tau$.

The test statistic is given by

$$WE = \frac{(\tau-1)^{-1}\sum_{i=1}^{\tau} w_i\,(\bar{z}_i - w^{-1}\sum_{i=1}^{\tau} w_i\,\bar{z}_i)^2}{1 + 2\,(\tau-2)(\tau^2-1)^{-1}\sum_{i=1}^{\tau}(n_i-1)^{-1}\,(1-w^{-1}w_i)^2}, \tag{5.2}$$

where $w_i = n_i\,/s_i^2, w = \sum_{i=1}^{\tau} w_i$ and $\bar{z}_i = n_1^{-1}\sum_j z_{ij}$. Under H_0, (5.2) follows an F-distribution with f_1 and f_2 degrees of freedom, with

$$f_1 = \tau - 1 \quad \text{and} \quad f_2 = \left[3\,(\tau^2-1)^{-1}\sum_{i=1}^{\tau}(n_i-1)^{-1}\left(1-\frac{w_i}{w}\right)^2\right]^{-1}.$$

In addition to the original WE test, we also consider a variant based on the ranks of $\{z_{ij}\}$. Since the ranks are assigned as for the KW test, that is over the entire observation period, this variant is referred to as the Welch-Kruskal-Wallis (WEKW) test.

Edwards' test

Tailored towards the analysis of epidemiological incidence data, Edwards' test assumes that the seasonality can be described by simple harmonic patterns Edwards (1958, 1961). Under the null hypothesis, the incidence rate is constant, i.e. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_\tau$. The test statistic is given by

$$ED = 8\left(\sum_{i=1}^{\tau}\sum_{j=1}^{J} z_{ij}\right)\frac{\{\sum_{i=1}^{\tau}\sqrt{\sum_{j=1}^{J} z_{ij}}\,\sin(2\pi i/\tau)\}^2 + \{\sum_{i=1}^{\tau}\sqrt{\sum_{j=1}^{J} z_{ij}}\,\cos(2\pi i/\tau)\}^2}{\sum_{i=1}^{\tau}\sqrt{\sum_{j=1}^{J} z_{ij}}} \tag{5.3}$$

The statistic follows a χ^2-distribution with two degrees of freedom. The null hypothesis is somewhat restrictive, as any kind of trend will be a deviation from a constant rate. Note also, that the test statistic entails the calculation of $\sqrt{\sum_{j=1}^{J} z_{ij}}$. Therefore, it can not be calculated if $\sum_{j=1}^{J} z_{ij} < 0$. Thus, it will usually not be possible to calculate Edwards' test based on some model's residuals.

Roger's test

The Roger (RO) test Roger (1977) is an enhancement of the ED test. H_0 is unchanged and the test statistic is

$$RO = \frac{2\{\sum_{i=1}^{\tau}\sum_{j=1}^{J} z_{ij}\sin(2\pi i/\tau)\}^2 + \{\sum_{i=1}^{\tau}\sum_{j=1}^{J} z_{ij}\cos(2\pi i/\tau)\}^2}{\sum_{i=1}^{\tau}\sum_{j=1}^{J} z_{ij}} \tag{5.4}$$

RO approximately follows a χ^2-distribution with two degrees of freedom.

Hewitt's test

The criterion devised by Hewitt et al. (1971) is a non-parametric alternative to Edward's test for monthly time series, i.e. $\tau = 12$. It assumes that a seasonal pattern results in an increased number of incidences in six adjacent months relative to the observations in all other months.[3]

Let $r_i = \text{rank}\left(\sum_{j=1}^{J} z_{ij}\right)$ and $M(i)$ be the set of the i-th month and its five consecutive months, eg $M(1) = \{1, 2, \ldots, 6\}$ and $M(10) = \{10, 11, 12, 1, 2, 3\}$. Then, the test statistic is given by

$$HE = \max\{q_1, \ldots, q_{12}\}, \tag{5.5}$$

where $q_i = \sum_{k \in M(i)} r_k$. Critical values have been derived by Walter (1980).

OCSB test

The OCSB test as developed by Osborn et al. (1988) is a seasonal unit root test, and thus is not designed to detect seasonal series in general, but only a specific manifestation of seasonality. Thus, the test does not try to distinguish seasonal from non-seasonal time series, but series with and without seasonal unit roots. It may still be worthwhile to include this test in an overall seasonality test that aims at being sensitive to various manifestations of seasonality. So, even if the OCSB test may not be the best indicator of seasonality per se, it may nonetheless contribute to an increase of the classification accuracy of an overall seasonality test.

The OCSB test follows the logic of the Dickey-Fuller test. If we assume a series to be integrated of non-seasonal and seasonal order I(1,D), we can model the dependence of y_t on y_{t-12} by

$$\nabla_1 y_t = \mu_t + \alpha \nabla_1 y_{t-12} + \epsilon_t, \tag{5.6}$$

where $\nabla_k = 1 - B^k$.

[3] A generalisation of Hewitt's test for seasonal pulses of less than six months has been developed by Rogerson (1996).

If $\alpha = 1$ there is a need for seasonal differencing. To see this, we can rewrite Equation 5.6 as follows

$$\begin{aligned}\nabla_1 y_t - \nabla_1 y_{t-12} &= \mu_t + \alpha_2 \nabla_1 y_{t-12} - \nabla_1 y_{t-12} + \epsilon_t \\ \nabla_1 \nabla_{12} y_t &= \mu_t + (\alpha_2 - 1)\nabla_1 y_{t-12} + \epsilon_t\end{aligned}$$

which for $\beta_2 = \alpha_2 - 1$ is equal to

$$\nabla_1 \nabla_{12} y_t = \mu_t + \beta_2 \nabla_1 y_{t-12} + \epsilon_t \tag{5.7}$$

Vice versa, if we assume a series to be integrated of non-seasonal and seasonal order I(d,1), we get

$$\nabla_{12} y_t = \mu_t + \alpha_1 \nabla_1 y_{t-12} + \epsilon_t \tag{5.8}$$

and after equivalent rearranging and redefining terms and $\beta_1 = \alpha_1 - 1$

$$\nabla_1 \nabla_{12} y_t = \mu_t + \beta_1 \nabla_{12} y_{t-1} + \epsilon_t \tag{5.9}$$

Adding Equation 5.7 and Equation 5.9 together and augmenting it by autoregressive terms up to order p to ensure white noise error terms, analogously to the reasoning of the augmented Dickey Fuller test (Said and Dickey, 1984), we can state the OCSB test regression as follows (cf. Rodrigues and Osborn, 1999):

$$\nabla_1 \nabla_{12} y_t = \mu_t + \beta_1 \nabla_{12} y_{t-1} + \beta_2 \nabla_1 y_{t-12} + \sum_{j=1}^{p} \phi_j \nabla_1 \nabla_{12} y_{t-j} + \epsilon_t \tag{5.10}$$

Here, we include three versions of the OCSB tests. The first OCSB test is augmented by non-seasonal autoregressive terms of order 3. The second OCSB (OCSB2) term additionally includes a seasonal autoregressive component of order one. The third version of the OCSB test (OCSB3) includes a non-seasonal autoregressive term of order 3 and seasonal autoregressive and moving average terms of order 1. By omitting seasonal autoregressive terms, as in the first version, we will not ensure white noise error terms, especially in the case of other forms of seasonality. Thus, we decrease the tests ability to detect seasonal unit roots correctly by letting it reject the null hypothesis more frequently in the case of other forms of seasonality, which may be desirable if the OCSB test is used as a seasonality test.

5.4 Variable selection and construction of overall test

Bearing in mind potential general conflicts between model parsimony and accuracy, we aim at condensing a set of candidate seasonality tests into an overall test that balances interpretable results against a low overall misclassification rate. Our approach essentially comprises two steps:

- the identification of the most informative candidate tests based on misclassification rates and VIMs in a CRF framework by iteratively eliminating the least informative test and
- the derivation of a classification rule based on an optimised single conditional inference tree which considers only the most informative candidate tests.

5.4.1 Identification of most informative seasonality tests

We use the following generic random feature elimination (RFE) algorithm to identify the most informative tests in a set of candidate seasonality tests:

1. Start with the entire set of candidate tests.
2. Train multiple CRFs on independent training data sets using the p-values of the current set of candidate tests as predictors and the seasonal status of simulated ARIMA processes as categorical response.
3. For each candidate test, average the VIM (5.1) over all CRFs. Store the mean and standard deviation of the overall misclassification rates computed for each CRF's external VAL data.
4. Eliminate the test with the smallest average VIM from the set of candidate tests.
5. Repeat Steps 2 to 4 until a pre-defined minimum number of predictors, p_{min}, is reached.
6. Select the most informative tests taking into account the overall misclassification rates along the entire RFE path and the numbers of tests included.

Remark 11. This RFE algorithm differs from similar approaches (Díaz-Uriarte and de Andrés, 2006; Genuer et al., 2010; Gregorutti et al., 2017) in three respects: first, we consider conditional instead of classical random forests (Step 2); second, we consider VIMs aggregated over multiple forests (Step 3) in order to lower the variance of the feature elimination in each recursion; third, we evaluate misclassification rates over external VAL instead of OOB data (Step 3).

5.4.2 Derivation of classification rule

Once the most informative seasonality tests have been identified in Step 6 of the generic RFE algorithm, their p-values are used again as predictors to grow a single conditional inference tree. Depending on the specified parameters, such as the number of identified tests, this tree is more or less likely to contain some redundancies in the sense that all branches after a particular node lead to the same classification. In this case, the single conditional inference tree is pruned to obtain a smaller but equivalent decision tree which always yields the same classification as the unpruned tree.

5.5 Simulation study

5.5.1 Design

We aim at simulating a set of seasonal and non-seasonal ARIMA models that closely mirrors the macroeconomic monthly time series stored in the Bundesbank's database in terms of their distributional properties. To this end, we first identify representative ARIMA models. Then, we use the simulation algorithm in Ollech and Webel (2020) that takes into account the shares of the identified ARIMA models as well as the relationship and distribution of the parameters within each identified class of ARIMA models. Finally, we apply the seasonality tests described in Section 5.3.1 to the simulated ARIMA models and construct the overall test as outlined in Section 5.4.

ARIMA model identification.

We randomly sample without replacement 3,300 macroeconomic time series that are seasonally adjusted each month and 10,600 monthly time series that are not seasonally adjusted at all from the database. For each sampled series, we run the automatic ARIMA model identification routines of the regARIMA and TRAMO pre-processors in JD+. For each identified model m of class $k \in \{\text{N-S}, \text{S}\}$, we calculate the model's averaged share p_{mk} among all models in class k and the simulation weights $w_{mk} = \tilde{p}_{mk} / \sum_m \tilde{p}_{mk}$, where $\tilde{p}_{mk} = p_{mk} \cdot \mathbb{1}_{\{p_{mk} \geq 0.01\}}$.

ARIMA model simulation.

For each identified model m of class $k \in \{\text{N-S}, \text{S}\}$ and each length $N \in \{60, 120, 240\}$, we simulate $100{,}000 \cdot w_{mk}$ ARIMA time series, using the same algorithm and parameter setup as in Ollech and Webel (2020) and the `{gsarima}` R package. Eventually, this yields a representative set of 600,000 simulated time series.

Candidate tests, RFE and CRF settings.

We start with a total of 18 seasonality tests as we consider three versions of the OCSB test: the OCSB1 test uses $J = 3$ in (5.10), the OCSB2 test essentially takes the sum over $j \in \{1, 2, 3, 12, 13\}$ in (5.10) and the OCSB3 test is the OCSB2 test with $\{\varepsilon_t\}$ being a seasonal MA(1) process instead of white noise. In general, the stationarity assumption of the candidate tests, if applicable, is met by differencing the input series once as in JD+.

We run the RFE algorithm with $p_{min} = 1$. In each recursion of Steps 2 to 4, 200 independent training data sets are randomly selected from the 600,000 simulated ARIMA models, where the size depends on the current number of candidate tests, varying between 800 and 8,000 due to computational restrictions. For each training data set, a CRF of $B = 100$ trees is grown with $n_{min} = 1$. In Step 3 a random sample of 50,000 ARIMA series not selected in the respective CRF are considered as external VAL data.

Regarding the overall test, we grow the single conditional inference tree on a training data set of 10,000 simulated ARIMA time series with $n_{min} = 50$. In our case, increasing the number of observations or lowering n_{min} leads to virtually the same tree except that additional nodes are grown that do not change the final classification.

5.6 Results for seasonality tests and random forests

Table 5.1 reports the misclassification rates of the candidate seasonality tests at the 1-percent significance level and CRFs[4] at the initial recursion of the RFE algorithm. The classification errors of the seasonality test obviously depend on the chosen significance level. In addition, the following observations can be made:

- The "-R" tests tend to have lower type I and type II misclassification rates than their standard implementations, except for the QS and RO tests.
- The error rates of the residual based tests tend to be less dependent on the length of the time series.
- The ED, RO and HE tests display unacceptably high misclassification rates, especially for seasonal series.[5] For that reason and then also to reduce the compational cost, they will not be considered for the RFE algorithm and construction of the overall seasonality test.
- For the OCSB tests, the order of augmentation strongly affects the misclassification rates: the OCSB1 test tends to have higher type I but lower type II misclassification

[4] The rates are based on the entire set of simulated ARIMA models for the candidate tests and averaged over the respective forests for the OOB and VAL data. Results for the 5-percent significance level can be found in Table 5.4.

[5] Originally, both tests have been tailored to incidence data in epidemiology and are not scale invariant as they depend strongly on the level of the input series.

Table 5.1: Misclassification rates as a percentage (N-S = non-seasonal, S = seasonal series), $\alpha = 0.01$.

		Simulated ARIMA series							
		All lengths		5-year		10-year		20-year	
Classifier		N-S	S	N-S	S	N-S	S	N-S	S
CRF	OOB	0.6	1.9	0.6	1.9	0.5	2.0	0.7	1.9
	VAL	0.6	1.9	0.6	2.0	0.5	1.8	0.6	1.8
Seasonality tests	QS	4.9	1.5	2.5	1.7	5.0	1.4	7.1	1.3
	QS-R	0.4	8.3	0.2	7.8	0.3	7.7	0.7	9.6
	FT	2.1	2.1	1.5	2.2	2.3	1.9	2.4	2.1
	FT-R	0.8	2.1	0.3	2.4	0.7	2.0	1.3	2.0
	KW	2.4	3.8	1.9	3.9	2.6	3.7	2.7	3.8
	KW-R	0.7	2.1	0.3	2.3	0.6	2.0	1.1	2.1
	PT	3.2	3.6	3.2	3.4	3.3	3.6	3.2	3.9
	SD	4.0	2.7	4.4	2.5	4.1	2.7	3.7	2.9
	RO	11.8	93.9	10.1	93.7	11.6	94.0	13.7	94.0
	RO-R	16.7	78.7	15.6	82.3	17.2	77.6	17.2	76.4
	ED	4.2	99.2	3.3	99.1	4.1	99.3	5.2	99.4
	HE*	12.0	94.3	11.8	96.5	12.0	94.7	12.3	91.7
	HE-R*	4.5	97.2	4.3	97.7	4.7	97.5	4.5	96.3
	WE	3.5	3.8	4.3	3.7	3.2	3.7	3.0	4.0
	WE-R	1.7	2.2	2.6	2.1	1.4	2.1	1.1	2.3
	WE-KW	5.6	3.4	9.0	3.1	4.5	3.4	3.4	3.7
	WE-KW-R	4.0	1.8	7.5	1.5	2.8	1.8	1.7	2.0
	OCSB	4.9	4.3	7.8	3.8	3.7	4.5	3.2	4.6
	OCSB2	3.3	4.9	8.4	3.1	1.0	5.0	0.6	6.6
	OCSB3	10.5	4.6	11.9	3.2	8.6	4.7	11.0	5.9

* For these tests the $\alpha = 0.05$ results are included, as none of the p-values are below 1 percent.

rates than the OCSB2 test, especially for longer series; the OCSB3 test does not outperform either of them in general.

The precision and negative classification values of the seasonality tests are included in Table 5.5. Basically, this information shows how often a test is correct whenever it states that a time series is non-seasonal resp. seasonal. The table adds the following insights:

- When residual based versions of the QS, FT and KW tests find a series to be seasonal, they are rarely incorrect.
- The modified QS test is particularly reliable when indicating non-seasonal behaviour.

All tests except for Roger's, Edwards' and Hewitt's tests are included in the conditional random forest based variable selection procedure. As discussed in section 5.4 in

Table 5.2: Precision and negative classification value (N-S = classified as non-seasonal, S = classified as seasonal), $\alpha = 0.01$.

Classifier		Simulated ARIMA series All lengths N-S	S	5-year N-S	S	10-year N-S	S	20-year N-S	S
CRF	OOB	98.1	99.4	98.1	99.4	98.0	99.5	98.1	99.2
	VAL	98.2	99.4	98.1	99.4	98.2	99.5	98.2	99.3
Seasonality tests	QS	98.5	95.3	98.2	97.5	98.5	95.2	98.6	93.2
	QS-R	92.3	99.6	92.8	99.8	92.9	99.7	91.3	99.3
	FT	97.9	97.9	97.8	98.5	98.1	97.7	97.9	97.6
	FT-R	97.9	99.2	97.7	99.7	98.0	99.3	98.0	98.7
	KW	96.3	97.6	96.2	98.1	96.4	97.4	96.3	97.2
	KW-R	97.9	99.3	97.7	99.7	98.0	99.3	98.0	98.9
	PT	96.4	96.8	96.6	96.8	96.4	96.7	96.1	96.8
	SD	97.3	96.0	97.5	95.7	97.3	95.9	97.1	96.3
	RO	48.6	33.8	49.1	38.1	48.6	34.0	48.0	30.1
	RO-R	51.6	55.9	50.8	53.0	51.8	56.5	52.2	57.7
	ED	49.3	15.0	49.6	21.9	49.3	14.9	49.0	10.2
	HE*	48.3	32.2	47.8	23.0	48.2	30.5	48.8	40.4
	HE-R*	49.6	38.4	49.5	34.2	49.5	34.6	49.7	45.0
	WE	96.2	96.5	96.3	95.7	96.3	96.8	96.1	97.0
	WE-R	97.9	98.3	97.9	97.4	97.9	98.6	97.7	98.9
	WE-KW	96.5	94.5	96.7	91.4	96.6	95.5	96.4	96.6
	WE-KW-R	98.2	96.1	98.4	92.9	98.2	97.2	98.0	98.3
	OCSB	95.7	95.1	96.1	92.4	95.6	96.2	95.5	96.8
	OCSB2	95.2	96.6	96.7	92.0	95.2	99.0	93.8	99.4
	OCSB3	95.1	90.1	96.5	89.0	95.1	91.7	93.7	89.5

* For these tests the $\alpha = 0.05$ results are included, as none of the p-values are below 1 percent.

each round 200 conditional random forests are grown and for each the variable importance is estimated. Afterwards, that variable is eliminated which has the lowest mean variable importance over all random forests.

5.6.1 RFE path

Table 5.3 reports the results of the RFE algorithm. The misclassification rate, averaged over external VAL data of 20 CRFs, gradually decreases from 1.18% to 1.04% during RFE recursions 1 to 13, then soars to 1.37% in the final recursion where only one candidate test is left. The decrease of the average misclassification rate reflects the elimination of tests that do not have a positive contribution to the reduction of the misclassification rate compared to the remaining tests. The minimum value of 1.04 % is achieved in both round 12 and 13. Taking additionally the respective standard deviations into account,

Table 5.3: Elimination of predictor in specified round of the conditional random forest based variable selection procedure.

Round	Eliminated test	Misclassification rate Average	Std.dev.
1	WE-KW	1.18	0.079
2	WE	1.16	0.080
3	PER	1.14	0.080
4	KW	1.13	0.086
5	SD	1.11	0.089
6	OCSB2	1.09	0.089
7	WE-KW-R	1.08	0.085
8	OCSB3	1.08	0.076
9	OCSB1	1.09	0.072
10	FT	1.08	0.068
11	WE-R	1.06	0.066
12	FT-R	1.04	0.051
13	QS	1.04	0.041
14	QS-R	1.08	0.038
15	KW-R	1.37	0.046

a set of three seasonality tests seems to provide a convenient balance between accuracy and parsimony and, thus, to be a good starting point for deriving the classification rule of the overall test.

The p-values of the KW-R, FT-R, WE-R and WE-KW-R are closely related, their pairwise Pearson correlation coefficient is at least 0.93 when computed over all simulated time series. This is especially true for the relationship between KW-R and FT-R, their correlation exceeds 0.96. To some extent, these tests can thus be viewed as substitutes, in the sense that they appear to be sensitive to a similar manifestation of seasonality. The correlation between these tests and the QS resp. the QS-R is in all cases below 0.74 resp. 0.77, while the correlation between QS and QS-R is 0.83. This indicates that an overall seasonality test composed of the QS, the QS-R and any of the ANOVA based tests should be able to identify various forms of seasonality, e.g. from very stable to continuously changing seasonal patterns.

5.7 Overall seasonality test

Based on Table 5.3 the KW-R, QS-R and QS test are identified as the most promising set of tests to function as building blocks for an overall seasonality test. Figure 5.2 shows a representative single unpruned conditional inference tree[6] based upon these three tests

[6]The exact result depends on the number of observations used to construct the tree and the minimum number of cases in a node to be considered for splitting. Yet, the relevant nodes remained virtually

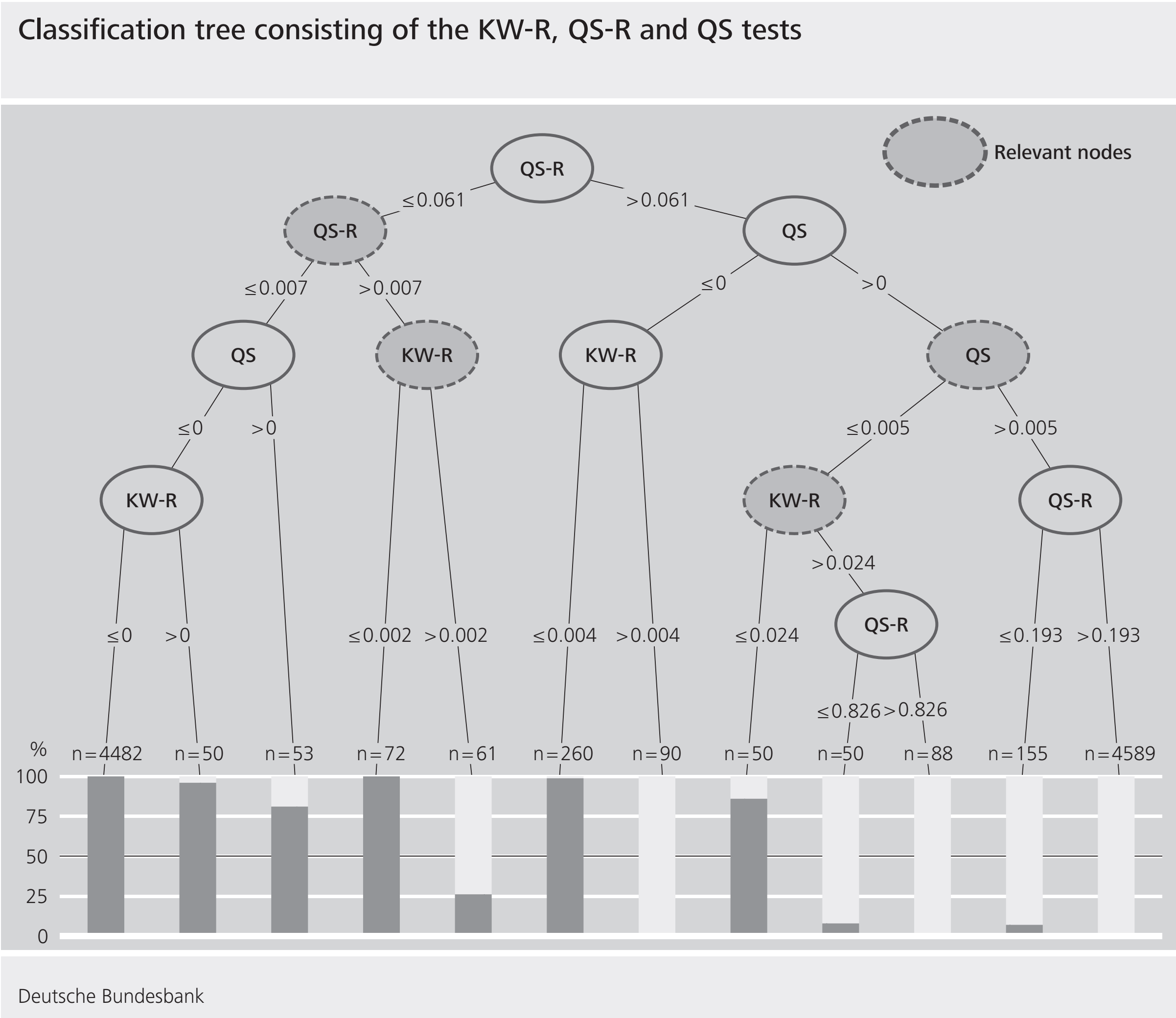

Figure 5.2: Classification tree consisting of KW-R, QS-R, and QS tests

and the classification of the training data, where all thresholds have been rounded to three decimal places. The tree actually consists of 11 nodes but due to some redundancies only three of them are relevant for the final classification. To see this, note, for example, that any series in the left daughter node of the initial node is classified as seasonal if p-value $\leq$ 0.007 holds for the QS-R test, regardless of all further branches. Also, the combination of the QS and KW-R test to the right of the initial node is encompassed by the subsequent combination of tests. The first combination of rules would find a series to be seasonal if the p-value ≤ 0 for the QS-test and p-value ≤ 0.004 for the KW-R test. This is always true, if the second combination of rules is true, i.e. if the p-value ≤ 0.005 for the QS-test and p-value ≤ 0.024 for the KW-R test. Therefore, the first combination of rules can be omitted, leading to the rule shown in algorithm 2.

The tree-based decision rule contains rather unconventional thresholds (p-values). Therefore, a simplified rule based on the same order of tests but with more conventional

unaffected by these parameters, instead the trees varied with respect to irrelevant splits, i.e. splits that do not affect the final classification.

Algorithm 2 Overall seasonality test: tree-based rule

if $p(\text{QS-R}) \le 0.01$ **then**
  Series is seasonal
else if $p(\text{KW-R}) \le 0.001$ **then**
  Series is seasonal
else if $p(\text{QS}) \le 0.01$ & $p(\text{KW-R}) \le 0.01$ **then**
  Series is seasonal
else
  Series is not seasonal
end if

thresholds may be warranted without compromising too much accuracy. To find such rule, grid search over pre-specified values and accuracy of the combination of tests is computed from training set of 100,000 observations. A rule of the structure of algorithm 2 reads:

1. p-value $\le c_1$ for the QS-R test
2. p-value $\le c_2$ for the KW-R test
3. p-value $\le c_3$ for the QS test & p-value $\le c_4$ for the KW-R test

We run the grid search over the thresholds $c_i \in \{0.001, 0.002, 0.01, 0.02, 0.05, 0.1\}$ for $i \in \{1, 2, 3, 4\}$. With $p(\cdot)$ denoting the p-value of a given seasonality test, the simplified tree-based rule is given in algorithm 3.

Algorithm 3 Overall seasonality test: simplified tree-based rule

if $p(\text{QS-R}) \le 0.01$ **then**
  Series is seasonal
else if $p(\text{KW-R}) \le 0.001$ **then**
  Series is seasonal
else if $p(\text{QS}) \le 0.01$ & $p(\text{KW-R}) \le 0.01$ **then**
  Series is seasonal
else
  Series is not seasonal
end if

Algorithm 4 Overall seasonality test: simplified logical rule

if $p(\text{QS-R}) \le 0.01$ **or**
$p(\text{KW-R}) \le 0.001$ **or**
$\{p(\text{QS}) \le 0.01$ **and** $p(\text{KW-R})\} \le 0.01$ **then**
  Series is seasonal
else
  Series is not seasonal
end if

Since the tree-based rule contains only three tests, the simplified tree-based rule can easily be translated into the following logical rule (see also algorithm 4):

1. p-value $\leq c_1$ for the QS-R test OR
2. p-value $\leq c_2$ for the KW-R test OR
3. p-value $\leq c_3$ for the QS test & p-value $\leq c_4$ for the KW-R test

The type I and type II misclassification rates of this classification rule are given by 0.59% and 1.47%, respectively, for the training data and by 0.55% and 1.52%, respectively, for the VAL data, that is the simulated ARIMA processes not used for optimising the classification rules.

Remark 12. This overall test is sensitive to both stable and moving seasonal pattern since the ANOVA-type rank test are constructed to identify relatively stable seasonal patterns, while the QS tests allows for a higher degree of time-varying seasonality.

5.7.1 Comparison to overall tests in JDemetra+

We can compare the accuracy of our combined seasonality test to the overall seasonality tests from X-13 and TramoSeats, both implemented in JDemetra+. The X-13 overall tests states whether seasonality is “present”. Based on this, the test achieves a misclassification rate of 6.9% over all series. As discussed above, TramoSeats differentiates between weak and strong seasonality. Using the “weak seasonality” indicator, the overall test incorrectly classifies 5.1% of all series. When considering only the series for which the test finds “strong seasonality” as seasonal and all other series as non-seasonal, the misclassification rate reduces to 4.5%.

This shows that none of the overall seasonality tests come close to the 1.3% that our combined seasonality test achieves.

5.8 Summary

We proposed a generic scheme to construct an overall seasonality test from a set of candidate tests with two building blocks: (1) recursive feature elimination in conditional random forests to identify the most informative candidate tests and (2) a pruned conditional inference tree based on the latter tests for the final classification. Using 600,000 ARIMA processes that are simulated to be representative of the Bundesbank’s time series database and 18 candidate tests, we identified residual-based variants of the QS and Kruskall-Wallis test as well as the traditional QS test as being most informative. The resulting classification rule is highly accurate and not overly complex.

Table 5.4: Misclassification rates as a percentage (N-S = non-seasonal, S = seasonal series).

Classifier			Simulated ARIMA series All lengths N-S	S	5-year N-S	S	10-year N-S	S	20-year N-S	S
CRF		OOB	0.6	1.9	0.6	1.9	0.5	2.0	0.7	1.9
		VAL	0.6	1.9	0.6	2.0	0.5	1.8	0.6	1.8
Seasonality tests	$\alpha = 0.05$	QS	7.4	1.2	4.9	1.4	7.5	1.1	9.8	1.2
		QS-R	1.3	6.9	1.0	6.3	1.1	6.3	1.7	8.1
		FT	6.6	1.6	5.7	1.6	7.0	1.5	7.2	1.7
		FT-R	4.1	1.6	3.0	1.7	4.1	1.5	5.1	1.6
		KW	6.9	3.2	6.2	3.1	7.1	3.1	7.4	3.3
		KW-R	3.9	1.6	2.9	1.6	4.1	1.5	4.8	1.7
		PT	8.1	3.2	8.3	2.8	8.2	3.1	8.0	3.5
		SD	9.1	2.2	9.7	2.0	9.2	2.2	8.5	2.5
		RO	15.1	91.0	13.5	90.7	14.9	91.1	16.8	91.3
		RO-R	19.9	71.7	18.8	74.9	20.6	70.6	20.4	69.8
		ED	4.9	99.1	3.9	98.9	4.8	99.1	6.1	99.3
		HE	12.0	94.3	11.8	96.5	12.0	94.7	12.3	91.7
		HE-R	4.5	97.2	4.3	97.7	4.7	97.5	4.5	96.3
		WE	8.9	3.3	10.8	3.0	8.2	3.2	7.7	3.6
		WE-R	6.3	1.7	8.4	1.5	5.6	1.7	5.0	2.0
		WE-KW	12.0	2.9	17.5	2.7	10.1	2.9	8.3	3.2
		WE-KW-R	10.3	1.4	16.5	1.2	8.3	1.4	6.1	1.6
		OCSB	2.9	7.1	4.0	7.0	2.4	7.4	2.3	6.7
		OCSB2	0.8	10.0	1.7	6.9	0.4	9.6	0.4	13.4
		OCSB3	3.5	8.5	3.6	7.0	3.1	8.6	4.0	9.8

5.A Appendix

5.A.1 Using the OCSB test as a pre-treatment

Instead of include the OCSB test as a seasonality test into the RFE path, another strategy is to use the OCSB test to identify time series with a seasonal unit root. Then, the elimination algorithm uses the remaining seasonality tests solely on the series without a seasonal unit root. In this setting, the OCSB test would not be re-purposed as a seasonality test, but serve its original function. The drawback is that, in this way, we do not estimate the informational content of the OCSB test. Furthermore, we do exempt the test from being included in a less apparent, possibly non-linear, manner in the nodes of the random forests.

To investigate the potential of the OCSB test as a pre-treatment, we use the OCSB1 test to identify seasonal unit roots. Keeping in mind that under the null hypothesis there

Table 5.5: Precision and negative classification value (N-S = classified as non-seasonal, S = classified as seasonal), $\alpha = 0.05$.

Classifier		Simulated ARIMA series All lengths N-S	S	5-year N-S	S	10-year N-S	S	20-year N-S	S
CRF	OOB	98.1	99.4	98.1	99.4	98.0	99.5	98.1	99.2
	VAL	98.2	99.4	98.1	99.4	98.2	99.5	98.2	99.3
Seasonality tests	QS	98.7	93.0	98.6	95.2	98.8	92.9	98.7	90.9
	QS-R	93.5	98.6	94.1	99.0	94.0	98.8	92.5	98.1
	FT	98.3	93.7	98.3	94.5	98.4	93.4	98.2	93.2
	FT-R	98.4	96.0	98.3	97.0	98.4	95.9	98.3	95.0
	KW	96.7	93.3	96.8	93.9	96.8	93.1	96.6	92.9
	KW-R	98.4	96.1	98.4	97.1	98.4	96.0	98.3	95.3
	PT	96.7	92.2	97.0	92.1	96.7	92.2	96.3	92.3
	SD	97.6	91.4	97.8	91.0	97.6	91.4	97.4	92.0
	RO	48.4	37.2	49.0	40.7	48.5	37.1	47.9	34.1
	RO-R	52.9	58.5	52.2	57.0	53.1	58.6	53.5	59.5
	ED	49.1	15.2	49.5	22.2	49.2	14.9	48.8	10.3
	HE	48.3	32.2	47.8	23.0	48.2	30.5	48.8	40.4
	HE-R	49.6	38.4	49.5	34.2	49.5	34.6	49.7	45.0
	WE	96.5	91.5	96.7	89.9	96.6	92.1	96.3	92.5
	WE-R	98.2	93.9	98.4	92.1	98.3	94.5	98.0	95.2
	WE-KW	96.8	89.0	96.9	84.6	96.9	90.5	96.6	92.1
	WE-KW-R	98.5	90.5	98.6	85.6	98.5	92.2	98.3	94.1
	OCSB	93.3	96.9	93.2	95.9	93.0	97.4	93.6	97.6
	OCSB2	90.9	99.1	93.4	98.2	91.3	99.5	88.2	99.6
	OCSB3	91.9	96.3	93.2	96.3	91.8	96.7	90.7	95.8

is a seasonal unit root, we use a high significance level of 10 percent to only include time series for which there is very low evidence of the absence of a seasonal unit root.

Table 5.6 shows that the misclassification rates cannot be further reduced by using the OCSB test beforehand.

Table 5.6: Elimination of predictor in specified round of the conditional random forest based variable selection procedure. With OCSB1 test used to detect series with seasonal unit roots before the random forests are used to detect remaining seasonal series.

		Mean variable importance	
		Misclassification rate	
Round	Test	Average	Std.dev.
1	WE-KW	1.38	0.111
2	WE	1.32	0.116
3	PER	1.31	0.119
4	KW	1.29	0.114
5	SD	1.25	0.097
6	WE-KW-R	1.22	0.095
7	FT	1.20	0.087
8	WE-R	1.20	0.082
9	FT-R	1.18	0.087
10	QS	1.16	0.067
11	QS-R	1.16	0.053
12	KW-R	1.44	0.041

Bibliography

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Díaz-Uriarte, R. and de Andrés, S. A. (2006). Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics*, 7:Article 3.

Edwards, J. (1958). Congenital Malformations of the Central Nervous System in Scotland. *British Journal of Preventive & Social Medicine*, 12(3):115–130.

Edwards, J. (1961). The recognition and estimation of cyclic trends. *Annals of Human Genetics*, 25(1):83–87.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable Selection Using Random Forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and Variable Importance in Random Forests. *Statistics and Computing*, 27(3):659–678.

Hewitt, D., Milner, J., Csima, A., and Pakula, A. (1971). On Edwards' Criterion of Seasonality and a Non-Parametric Alternative. *British journal of preventive & social medicine*, 25(3):174–176.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(3):1–22.

Maravall, A. (2011). *Seasonality Tests and Automatic Model Identification in TRAMO-SEATS*. Bank of Spain. Mimeo.

Morry, M. and Lothian, J. (1978). A Test for the Presence of Identifiable Seasonality When Using the X-11-ARIMA Program. Research Paper 78-10-002, Seasonal Adjustment and Time Series Analysis Staff, Statistics Canada.

Ollech, D. and Webel, K. (2020). A random forest-based approach to identifying the most informative seasonality tests. Discussion Paper No 55/2020, Deutsche Bundesbank.

Osborn, D. R., Chui, A. P., Smith, J. P., and Birchenhall, C. R. (1988). Seasonality and the Order of Integration for Consumption. *Oxford Bulletin of Economics and Statistics*, 50(4):361–377.

Rodrigues, P. M. M. and Osborn, D. R. (1999). Performance of Seasonal Unit Root Tests for Monthly Data. *Journal of Applied Statistics*, 26(8):985–1004.

Roger, J. (1977). A Significance Test for Cyclic Trends in Incidence Data. *Biometrika*, 64(1):152–155.

Rogerson, P. A. (1996). A Generalization of Hewitt's Test for Seasonality. *International Journal of Epidemiology*, 25(3):644–648.

Said, S. E. and Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, 71(3):599–607.

Strasser, H. and Weber, C. (1999). On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics*, 8:220–250.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9:Article 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8:Article 25.

Walter, S. (1980). Exact Significance Levels for Hewitt's Test for Seasonality. *Journal of Epidemiology and Community Health*, 34(2):147–149.

Webel, K. and Ollech, D. (2018). An overall seasonality test based on recursive feature elimination in conditional random forests. In *Proceedings of the 5th International Conference on Time Series and Forecasting*, pages 20–31. Granada.

Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika*, 38(3/4):330–336.

CHAPTER 6

Unit Root Test Combination via Random Forests

LUCA NOCCIOLA, DANIEL OLLECH AND KARSTEN WEBEL

Abstract

There is a wide variety of non-seasonal and seasonal unit root tests. However, it is not always obvious which tests can be relied upon due to uncertainties in identifying the data generating process, often with respect to the presence of deterministic terms and the initial conditions. We evaluate the size and power of a large set of unit root tests on time series that are simulated to be representative of economic time series in the M4 competition data. Furthermore, using a conditional random forest-based elimination algorithm, we assess which tests should be combined to improve the performance of each individual test.

Published as: Nocciola, L., Ollech, D., and Webel, K. (2021). Unit Root Test Combination via Random Forests. In *International Conference on Time Series and Forecasting* (pp. 31–46). Cham: Springer International Publishing.

Author contributions: D.O. designed the simulation study and random feature elimination algorithm. D.O. and L.N. implemented all algorithms and previously unavailable unit root tests. D.O. and L.N. analysed the results. L.N. and K.W. wrote the manuscript. D.O reviewed the manuscript.

Keywords: ARIMA time series; conditional inference trees; Monte Carlo simulation; sequential testing; supervised machine learning

6.1 Introduction

The question of whether a given macroeconomic time series contains a unit root is important since the presence (absence) of such a root implies that exogenous shocks have a persistent (transient) effect on the data. It is also well-known that unit roots cause problems in inference and forecasting. First, we know that, for independent random walks, we cannot rely on sample correlations and regression coefficients as consistent measures of the relationships within the population, since they can spuriously take on any value within the interval $[-1, 1]$ and $\mathbb{R}$, respectively, even asymptotically. Second, the rejection rate of standard t-tests for the null hypothesis of a unit root increases with sample size. Accordingly, these tests are not helpful due to the fact that t-statistics do not converge to any asymptotic distribution unless properly normalised at a customised rate (Ernst et al., 2017; Granger and Newbold, 1974; Hassler and Hosseinkouchack, 2022; Phillips, 1986; Yule, 1926). Third, forecast accuracy under unit roots in the autoregressive polynomial steadily deteriorates as the horizon increases, since the forecast error variance grows linearly with the forecast horizon and tends towards infinity in the limit, contrary to the stationary case (Clements and Hendry, 2001).

The large number of competing non-seasonal and seasonal unit root tests can complicate the researcher's decision regarding which test to apply in a given situation, even if the seasonal status of the data is known, or suggested by a seasonality test. Size distortions and power issues have been reported for various scenarios of practical importance (Wolters and Hassler, 2006), including finite samples (Cochrane, 1991), near-unit root behaviour, and particularly large negative roots in the moving average polynomial (Ng and Perron, 2001; Rodrigues and Osborn, 1999). Contradictory conclusions have also been repeatedly reported when several tests are applied to the same data, especially for seasonal unit root tests as a result of the different model specifications that they impose (Rodrigues and Osborn, 1999). Uncertainty surrounding the deterministic mean function and the initial conditions is also known to affect the tests' performance (Elliott and Müller, 2006; Müller and Elliott, 2003; West, 1987) and, moreover, make the assumptions of standard testing procedures unlikely to be met. Various strategies for dealing with these types of uncertainty have been suggested for non-seasonal unit root tests, including sequential pre-testing for trend specifications and unit roots, data-dependent weighted averaging of unit root tests and running union-of-rejections decision rules (Ayat and Burridge, 2000; Harvey et al., 2009; Perron, 1988). Although these strategies may involve multiple tests, they usually work with only a small pre-selection of tests, often from the same family. We elaborate on this approach by reinterpreting the unit root hypothesis as a classification problem with two classes. Our sequential approach utilises the conditional random forest classifier to identify, rank and combine the most informative tests. It is thus capable of incorporating a larger set of unit root tests and naturally extends to seasonal unit root tests. Our approach can help practitioners to select unit root tests with lower misclassification rates.

The remainder of this paper is organised as follows: Section 6.2 reviews widely applied tests for non-seasonal and seasonal unit roots, Section 6.3 provides basic information about random forests, Section 6.4 explains our proposed testing strategy, Section 6.5 reports our results, and, finally, Section 6.6 summarises.

6.2 Unit Root Tests

Let $\{y_t\}$ be a discrete time series with τ observations per year and assume that the series can be adequately described by the model

$$\phi(B)\,(y_t - \mu_t) = \theta(B)\,\varepsilon_t \quad , \tag{6.1}$$

where B is the backshift operator, $B^k x_t = x_{t-k}$, $\phi(B)$ is an autoregressive (AR) polynomial that has a root on or outside the unit circle, $\theta(B)$ is a moving average (MA) polynomial that has roots outside the unit circle, $\{\mu_t\}$ is a deterministic function of time, and $\{\varepsilon_t\}$ is a zero-mean white noise process with finite variance.

The series $\{y_t\}$ is said to be integrated of order $d \in \mathbb{N}_0$, denoted by $\{y_t\} \sim I(d)$, if $\phi(B)$ contains the factor $(1-B)^d$. Similarly, the series is said to be seasonally integrated of order $D \in \mathbb{N}_0$, denoted by $\{y_t\} \sim SI(D)$, if $\phi(B)$ contains the factor $(1-B^\tau)^D$. Thus, $\{y_t\}$ is stationary around $\{\mu_t\}$ if $(d, D) = (0, 0)$, non-stationary if $(d, D) \neq (0, 0)$, and invertible in either case.

6.2.1 Non-Seasonal Unit Roots

The problem of testing for a non-seasonal unit root can be stated as

$$\mathcal{H}_0 : \{y_t\} \sim I(1) \text{ versus } \mathcal{H}_1 : \{y_t\} \sim I(0) \tag{6.2}$$

and several tests have been suggested under different assumptions for the AR and MA polynomials in (6.1), with the common assumption being that there is no seasonality, or at least no seasonal unit root ($D = 0$), in the data. The function $\{\mu_t\}$ is usually represented by

$$\mu_t = \mu + \beta t \quad , \tag{6.3}$$

covering absence of deterministic terms ($\mu = \beta = 0$) as well as deviations of $\{y_t\}$ from a constant mean ($\beta = 0$), from a linear trend with a zero mean ($\mu = 0$), and from a linear trend with a non-zero mean (unrestricted model).

The Dickey-Fuller (DF) test (Dickey and Fuller, 1979, 1981) considers the pure first-order AR case, i.e. $\phi(B) = 1 - \rho B$ and $\theta(B) = 1$, so that (6.2) is reduced to testing $\rho = 1$ against $\rho < 1$. The DF test is based on the time series regression

$$y_t = \mu_t + \rho y_{t-1} + \eta_t \quad , \tag{6.4}$$

where $\{\eta_t\}$ coincides with $\{\varepsilon_t\}$ of (6.1) under Gaussianity and the initial condition is $y_0 = 0$. The two proposed one-sided tests are the conventional t-statistic $\mathrm{DF}_\tau = (\hat{\rho} - 1) \times \hat{\sigma}_{\hat{\rho}}^{-1}$ and $\mathrm{DF}_\rho = T\,(\hat{\rho} - 1)$, where T is the sample size and estimates are obtained by ordinary least squares (OLS). However, the null distribution of either statistic depends on the form of $\{\mu_t\}$ and is non-standard in each case, so that revised sets of critical values apply (Dickey and Fuller, 1981; Fuller, 2009).

Higher-order AR and ARMA models can be dealt with by the augmented Dickey-Fuller (ADF) test Said and Dickey (1984), even for unknown model orders. Letting $\Delta = 1 - B$, this test is usually carried out by testing $\rho^\star = 0$ against $\rho^\star < 0$ in the augmented regression

$$\Delta\, y_t = \mu_t + \rho^\star y_{t-1} + \sum_{j=1}^{k-1} \phi_j^\star \Delta\, y_{t-j} + \varepsilon_t \quad , \tag{6.5}$$

using the OLS t-statistic $\mathrm{ADF}_\tau = \hat{\rho}^\star \times \hat{\sigma}_{\hat{\rho}^\star}^{-1}$. If the lag length satisfies $k = p$ for pure AR(p) models and $k \to \infty$ at a controlled rate as $T \to \infty$ for ARMA models, then the DF critical values apply. Counterparts to critical F-values for testing against specific trend alternatives have also been tabulated (Dickey and Fuller, 1981).

Allowing for correlated and possibly heterogeneously distributed innovations, the Phillips-Perron (PP) statistics PP_τ and PP_ρ (Phillips, 1987; Phillips and Perron, 1988) rely on non-parametric transformations rather than augmentation to correct the DF_τ and DF_ρ statistics for the effects of nuisance parameters associated with the distribution of $\{\eta_t\}$. The PP statistics also follow the DF limiting null distributions and thus the DF critical values apply for each form of $\{\mu_t\}$.

Several extensions of the (A)DF and PP tests have been suggested, especially with respect to the proper treatment of the uncertainty surrounding $\{\mu_t\}$. For example, the Zivot-Andrews (ZA) test (Zivot and Andrews, 2002) allows for the estimation of break-points in $\{\mu_t\}$ under the alternative hypothesis, and the Elliott-Rothenberg-Stock (ERS) tests (Elliott et al., 1996) are essentially feasible asymptotically point-optimal DF_ρ tests of $\rho = \bar{\rho}$, allowing for polynomial trends. The test statistic is $\mathrm{ERS}_\rho = [S(\bar{\rho}) - \bar{\rho} S(1)] \times \hat{\omega}^{-2}$, where $S(\alpha)$ is the sum of the squared quasi-α-detrended series $\left\{ y_t - \hat{\mu}_t^{(\alpha)} \right\}$ and $\hat{\omega}^2$ is a consistent estimator of the sum of the covariances of $\{\eta_t\}$. A special member of the ERS family is a modified ADF_τ test that results from a specific choice of $\bar{\rho}^\star$ given T and nominal size. This test, denoted by ERS_τ, is essentially the ADF_τ test applied to the quasi-$\bar{\rho}^\star$-detrended series in (6.5) without $\{\mu_t\}$.

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski et al., 1992) considers the problem of testing for (trend) stationarity against non-stationarity, i.e.

$$\mathcal{H}_0 : \{y_t\} \sim I(0) \text{ versus } \mathcal{H}_1 : \{y_t\} \sim I(1) \ ,$$

assuming that $\{y_t\}$ is decomposable into a deterministic trend, random walk and stationary error. The test regression is essentially model (6.4) with $\rho = 0$ and $\{\mu_t\}$ being replaced with a random walk $\{\tilde{\mu}_t\}$:

$$y_t = \tilde{\mu}_t + \beta t + \eta_t \text{ with } \tilde{\mu}_t = \tilde{\mu}_{t-1} + \kappa_t \ , \tag{6.6}$$

where $\{\kappa_t\}$ is white noise with a zero mean and finite variance σ_κ^2. The KPSS test is the one-sided upper-tail Lagrange multiplier (and also the locally best invariant) test of $\sigma_\kappa^2 = 0$ against $\sigma_\kappa^2 > 0$, i.e. $\text{KPSS} = \hat{\sigma}_\eta^{-2} \sum_{t=1}^T S_t^2$, where $\{S_t\}$ is the partial-sum process of the residuals estimated from (6.6) under the null hypothesis and $\hat{\sigma}_\eta^2$ is the sum of the squared residuals divided by T. The asymptotic distribution of the KPSS statistic had been derived initially under the assumption that $\{\eta_t\}$ is Gaussian white noise with finite variance but has then been shown to also hold under weaker (strong mixing) regularity conditions of the PP tests (Phillips and Perron, 1988).

As an alternative to classical unit root tests, the Gómez-Maravall (GM) algorithm developed in Gómez and Maravall (2001) determines the appropriate differencing order for ARIMA models by iteratively analysing the roots in the AR and MA polynomials of different pre-defined seasonal ARMA models. Any root of the characteristic polynomial is defined to indicate a unit root if its modulus is larger than a pre-defined threshold that, by default, depends on the ARMA model under consideration.

6.2.2 Seasonal Unit Roots

In analogy to (6.2), several tests for the general problem

$$\mathcal{H}_0 : \{y_t\} \sim SI(1) \text{ versus } \mathcal{H}_1 : \{y_t\} \sim SI(0) \tag{6.7}$$

exist, where the deterministic component $\{\mu_t\}$ may now also contain seasonal dummies in (6.1), so that (6.3) is extended to

$$\mu_t = \mu + \beta t + \boldsymbol{\gamma}^\top \mathbf{d}_t \ , \tag{6.8}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_\tau)^\top$ is the vector of the τ fixed seasonal effects and $\mathbf{d}_t = (d_{1t}, \ldots, d_{\tau t})^\top$ with $d_{it} = 1$ if t falls within the i-th month or quarter and $d_{it} = 0$ otherwise.

The Dickey-Hasza-Fuller (DHF) test (Dickey et al., 1984) is essentially a seasonal variant of the DF test as it considers the pure first-order seasonal AR case, i.e. $\phi(B) = 1 - \rho_\tau B^\tau$ and $\theta(B) = 1$, so that (6.7) is reduced to testing $\rho_\tau = 1$ against $\rho_\tau < 1$ in the time series regression

$$y_t = \mu_t + \rho_\tau y_{t-\tau} + \eta_t \ ,$$

where $\{\mu_t\}$ is now given by (6.8). The two proposed one-sided tests are based on the conventional t-statistic $\mathrm{DHF}_\tau = (\hat{\rho}_\tau - 1) \times \hat{\sigma}_{\hat{\rho}_\tau}^{-1}$ and on $\mathrm{DHF}_\rho = T\,(\hat{\rho}_\tau - 1)$. OLS and symmetric least squares variants (including revised sets of critical values) are provided for the zero-mean ($\mu = \beta = 0$ and $\boldsymbol{\gamma} = \mathbf{0}$), constant-mean ($\mu \neq 0$, $\beta = 0$ and $\boldsymbol{\gamma} = \mathbf{0}$), and seasonal-means cases ($\mu = \beta = 0$ and $\boldsymbol{\gamma} \neq \mathbf{0}$). An ADF-like extension, which is referred to as the augmented DHF (ADHF) test, can be obtained via augmentation with lags of $\{\Delta_\tau y_t\}$ as in (6.5), where $\Delta_\tau = 1 - B^\tau$.

The Osborn-Chui-Smith-Birchenhall (OCSB) test (Osborn et al., 1988; Rodrigues and Osborn, 1999) considers $\phi(B) = (1 - \rho B)(1 - \rho_\tau B^\tau)$ and tests $(\rho, \rho_\tau) = (1, 1)$ against $(\rho, \rho_\tau) \neq (1, 1)$ in the (re-parameterised) regression

$$\Delta\Delta_\tau y_t = \mu_t + \beta_1 \Delta_\tau y_{t-1} + \beta_2 \Delta\, y_{t-\tau} + \eta_t \tag{6.9}$$

with the F-test for $(\beta_1, \beta_2) = (0, 0)$. However, model (6.9) also allows for sequential t-tests against one-sided (stationary) alternatives. If $\beta_2 = 0$, then the t-test for $\beta_1 = 0$ is the (A)DF test for the need of Δ in addition to Δ_τ. Similarly, if $\beta_1 = 0$, then the t-test for $\beta_2 = 0$ is the DHF test for the need of Δ_τ in addition to Δ. Revised critical values apply in either case. Moreover, assuming the validity of Δ, (6.9) can be rewritten as

$$\Delta_\tau z_t = \mu_t + \beta_1\, S(B)\, z_{t-1} + \beta_2\, z_{t-\tau} + \eta_t \ ,$$

where $\{z_t\} = \{\Delta\, y_t\}$ and $S(B) = 1 + B + \cdots + B^{\tau-1}$ is the annual aggregation operator, allowing for separate treatments of the roots in $\Delta_\tau = \Delta\, S(B)$.

The Hylleberg-Engle-Granger-Yoo (HEGY) test (Beaulieu and Miron, 1993; Franses, 1991; Hylleberg et al., 1990) extends this factorisation principle even further as it expands $\phi(B)$ about all roots of Δ_τ, additionally allowing for individual assessments (and different moduli) of the $\tau - 1$ unit roots of $S(B)$. The test regression reads

$$\Delta_\tau y_t = \mu_t + \sum_{i=1}^{\tau} \pi_i\, x_{i,t-1} + \eta_t \ ,$$

where the processes $\{x_{i,t-1}\}$ are non-singular linear transformations of lagged versions of $\{y_t\}$. The DHF null hypothesis thus implies that $\pi_i = 0$ for all $i \in \{1, \ldots, \tau\}$ but specific sub-hypotheses for single (real-valued or pairs of complex-valued) unit roots in Δ_τ can also be tested with separate t-tests and F-tests, respectively. Depending on the specified alternative and the form of $\{\mu_t\}$, the DF and DHF critical values apply except for $\boldsymbol{\gamma} \neq \mathbf{0}$, where critical values are found through simulations by Hylleberg et al. (1990).

The GM algorithm discussed above can also be used to determine the seasonal differencing order.

6.3 Random Forests

A random forest (RF) as described by Breiman (2001) is a supervised machine learning algorithm. It is a collection of classification trees constructed on bootstrap samples of the original training data. Additionally, at each split, the trees are restricted in that only a sub-sample of the available predictors is taken into consideration to determine the optimal split. A conditional RF as developed by Hothorn et al. (2006) enhances the classical RF in the presence of correlated predictors.

6.3.1 Classical Random Forests

Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_P)$ be a set of predictors with $\mathbf{x}_j = (x_{1j}, \dots, x_{Nj})^\top$ for all $j \in \{1, \dots, P\}$ and $\mathbf{y} = (y_1, \dots, y_N)^\top$ be a vector of N responses with $y_i \in \{0, \dots, k\}$ for all $i \in \{1, \dots, N\}$, constituting our training set $\mathcal{L}$.[1] Let $\mathcal{L}_b$ be a bootstrap sample of size N drawn from $\mathcal{L}$, from which we grow an unpruned classification tree $\mathcal{T}_b$ with M terminal nodes corresponding to M classification regions. To create a binary split of any terminal node m, we draw a random sample $\tilde{\mathbf{X}}$ of size $\tilde{P} < P$ without replacement from $\mathbf{X}$. For each sampled predictor $\tilde{\mathbf{x}}_j$, we determine the best split of node m amongst all possible splits of m, and the predictor that generates it (say $\tilde{\mathbf{x}}_{j*}$) is chosen as the splitting predictor for m. The optimal split is identified by, for example, the Gini index. Let $\hat{q}_{mk} = N_m^{-1} \sum_{\mathbf{x}_i \in R_m} \mathbb{I}\{y_i = k\}$ be the share of training data in node m from class k, where $N_m = \sum_i \mathbb{I}\{\mathbf{x}_i \in R_m\}$ and $\mathbf{x}_i = (x_{i1}, \dots, x_{iP})$ and R_m denote the i-th observation of the P predictors and the classification region corresponding to m, respectively. The Gini index is given by

$$Q_m(\mathcal{T}_b) = \sum_k \hat{q}_{mk}(1 - \hat{q}_{mk}) \ .$$

We stop the trees from growing whenever a pre-specified minimum number of observations in the terminal nodes is reached (say $N_{\min}$) or node impurity cannot be decreased further. Classification in the RF is obtained by an unweighted mode of the tree classifications. Moreover, we can determine predictor importance in an RF, for example, via the mean decrease in prediction accuracy after randomly permuting the values of the predictor $\mathbf{x}_j$ in the out-of-bag samples (i.e. the training data not boostrapped in $\mathcal{L}_b$), denoted by $\mathcal{O}_b$,

[1] In our case, y_i is binary ($k = 1$). However, $k = 3$ could also be used if we considered the combination of possibilities of having seasonal and non-seasonal unit roots.

to mimic the absence of $\mathbf{x}_j$. Let $\hat{y}_i(\mathcal{T}_b, \mathbf{x}_j)$ and $\hat{y}_i(\mathcal{T}_b, \mathbf{x}_{\pi(j)})$ be the predicted classes of y_i obtained from $\mathcal{T}_b$ before and after random permutation of the values of $\mathbf{x}_j$ in $\mathcal{O}_b$, where $\pi(\cdot)$ is the permutation scheme. The predictor importance of $\mathbf{x}_j$ is then given by

$$PI(\mathbf{x}_j) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i \in \mathcal{O}_b} \left[\frac{\mathbb{I}\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{x}_{\pi(j)})\}}{|\mathcal{O}_b|} - \frac{\mathbb{I}\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{x}_j)\}}{|\mathcal{O}_b|} \right] ,$$

where B is the number of classification trees in the forest.

6.3.2 Conditional Random Forests

The conditional RF introduces two adaptations with respect to predictor selection and the predictor importance measure. Predictors that are not related to $\mathbf{y}$ (identified via a test of independence between $\mathbf{y}$ and $\tilde{\mathbf{x}}_j$) are excluded in advance, while, among the predictors that exhibit association with $\mathbf{y}$ (via the same test), the predictor with the strongest association with $\mathbf{y}$, say $\tilde{\mathbf{x}}_{j*}$, is chosen as a splitting predictor used to determine the optimal split of node m. Let each node m be represented by a vector of integer case weights $\mathbf{w}_m = (w_{m,1}, \ldots, w_{m,N})^\top$, where $w_{m,i} > 0$ if $(\mathbf{x}_i,\, y_i)$ belongs to m and equals zero otherwise. Formally, the global null hypothesis can be formulated as

$$\mathcal{H}_0^{(m)} : \bigcap_{j=1}^{\tilde{P}} \mathcal{H}_0^{(m,j)} \text{ with } \mathcal{H}_0^{(m,j)} : \mathcal{D}(\mathbf{y}|\tilde{\mathbf{x}}_j, \mathbf{w}_m) = \mathcal{D}(\mathbf{y}|\mathbf{w}_m) \ ,$$

where $\mathcal{D}(\cdot)$ denotes an arbitrary distribution. The rejection rule of this hypothesis is based on the minimal (adjusted) p-value for rejecting the local hypothesis $\mathcal{H}_0^{(m,j)}$. The predictor $\tilde{\mathbf{x}}_{j*}$ can be selected from the local null hypothesis $\mathcal{H}_0^{(m,j*)}$ rejected at the smallest (adjusted) p-value. Finally, $\tilde{\mathbf{x}}_{j*}$ is used as a splitting predictor to find the best binary split of m according to a pre-specified splitting criterion, after which the case weights $\mathbf{w}_{m_L}$ and $\mathbf{w}_{m_R}$ of the left and right descendents of m are computed.

To determine predictor importance, the conditional RF uses a conditional permutation taking into account correlations among predictors, thus preventing seemingly influential predictors (due to their correlation with the truly influential predictors) from being attached high importance. The random permutation $\pi(\cdot)$ is now applied to the values of $\mathbf{x}_j$ only within subgroups of observations, say $\mathbf{x}_j^C$.[2] The conditional permutation-based predictor importance is given by

$$PI^C(\mathbf{x}_j) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i \in \mathcal{O}_b} \left[\frac{\mathbb{I}\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{x}_{\pi(j)|\mathbf{x}_j^C})\}}{|\mathcal{O}_b|} - \frac{\mathbb{I}\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{x}_j)\}}{|\mathcal{O}_b|} \right] , \tag{6.10}$$

[2] Alternatively, though less conservative, we can condition only on those predictors $\mathbf{x}_j^C$ whose correlation with $\mathbf{x}_j$ exceeds a certain threshold.

where $\pi(\cdot)|\mathbf{x}_j^C$ is the conditional permutation scheme and, for each tree $\mathcal{T}_b$, the permutation grid for $\mathbf{x}_j$ is given by the cut-points of $\mathbf{x}_j^C$ in $\mathcal{T}_b$.

6.4 Test Evaluation

We aim to evaluate the unit roots tests on ARIMA time series that are representative of the data used in the M4 competition (Makridakis et al., 2018, 2020). To this end, we simulate correlated Gaussian ARMA parameters and transform them so that their distribution is similar to the empirical distribution estimated from the M4 competition data Ollech and Webel (2020). This is achieved by combining the "NORmal-To-Anything" (NORTA) algorithm from Cario and Nelson (1997) with logspline density estimation from Stone et al. (1997). We simulate ARMA parameters and use these to simulate 120,015 time series. The set consists of 40,005 time series with 5, 10 and 20 years of monthly data, respectively. The set of time series is balanced in the sense that half of them contain a (seasonal) unit root. We then run the unit root tests on the raw and first-differenced simulated data and use their p-values as predictors in the conditional RF in order to classify the simulated data into time series with and without unit roots.[3] Finally, we use (6.10) to quantify the importance of each individual test.

Based on the test results, we run a conditional RF-based recursive feature elimination (CRFE) algorithm as developed by Webel and Ollech (2018). The key idea is to start with the full set of candidate tests and to grow a sequence of conditional RF by (1) eliminating the least important test in each round and (2) re-estimating the conditional RF with the reduced set of tests in the next round. This scheme is repeated until an "optimal" set of tests is obtained. In each round, we use the conditional RF to classify a test set of simulated time series not considered in tree growing. By evaluating the means and standard deviations of the misclassification rates in each round, we can assess the performance of the respective set of tests and eventually identify the "optimal" set.

6.5 Results

Table 6.1 reports the rejection rates for the non-seasonal unit root tests, using a nominal 1% significance level.[4] The ERS and GM tests have acceptable size properties but the former have relatively low power. The ADF, KPSS and PP tests display severe size distortions but compensate those with a relatively high power (above 70% except for the

[3]For computational reasons and parsimony, the tests have been performed with minimal lag length. Seasonal differences and the seasonal status of the series are not considered. The finite-sample critical values that we obtained from our own simulations in R (version 3.5.1) are very similar to the ones tabulated in the original papers (details are available upon request).

[4]Similar results are obtained for a nominal 5% significance level.

ADF test without deterministic terms). The ZA test has the worst size properties but the highest power among all tests considered. Overall, size issues seem to increase with the length of series for most tests, whereas the power tends to improve. Similar results are reported in DeJong et al. (1992b,a); Diebold and Rudebusch (1991); Schwert (1989). In particular, DeJong et al. (1992b) do not recommend using unit root tests for series with less than 100 observations.

Table 6.2 contains the rejection rates for the seasonal unit root tests, where we restrict ourselves to the use of the joint F-test for all roots for the HEGY tests and to the t-statistic for the seasonal unit root for the OCSB tests. All tests show severe size distortions except for the GM test when being applied to the raw time series. The HEGY tests are especially biased as they reject a true unit root null hypothesis in almost all cases, in particular for the longer series. On the contrary, almost all tests display an acceptable power of more than 80%, especially for the 10- and 20-year series, with the exception of the ADHF test with seasonal dummies when being applied to the raw data.

Table 6.1: Rejection rates for non-seasonal unit root tests (%, $\alpha = 0.01$)

		$d=0$				$d=1$			
Test	Model (6.3)	All	5-year	10-year	20-year	All	5-year	10-year	20-year
ADF_τ	$\mu = 0, \beta = 0$	25.5	17.2	23.2	36.2	19.0	17.0	19.1	20.9
	$\mu \neq 0, \beta = 0$	76.6	73.7	77.5	78.6	23.4	21.2	24.4	24.6
	$\mu \neq 0, \beta \neq 0$	88.9	82.1	92.8	91.9	32.7	27.8	35.4	34.9
ERS_τ	$\mu \neq 0, \beta = 0$	31.6	19.3	32.5	43.1	3.3	2.3	3.5	4.1
	$\mu \neq 0, \beta \neq 0$	40.4	17.3	44.2	59.7	8.1	5.6	8.8	9.8
ERS_ρ	$\mu \neq 0, \beta = 0$	49.4	41.7	49.4	57.0	6.2	8.2	6.0	4.5
	$\mu \neq 0, \beta \neq 0$	46.5	36.7	45.6	57.3	9.3	11.1	8.7	8.1
GM	$\mu \neq 0, \beta = 0$	77.9	75.9	78.2	79.6	9.6	17.0	8.7	3.2
PP_ρ	$\mu \neq 0, \beta = 0$	85.0	87.4	84.9	82.8	15.3	15.7	16.0	14.4
	$\mu \neq 0, \beta \neq 0$	97.0	96.5	98.0	96.6	27.5	26.1	29.1	27.2
PP_τ	$\mu \neq 0, \beta = 0$	85.9	87.9	85.8	84.1	28.5	27.8	29.4	28.3
	$\mu \neq 0, \beta \neq 0$	97.3	96.7	98.3	97.0	38.6	36.6	40.5	38.7
ZA	$\mu \neq 0, \beta = 0$	97.2	94.9	98.2	98.5	42.7	36.1	44.8	47.3
	$\mu = 0, \beta \neq 0$	97.4	94.7	98.3	99.0	45.6	37.6	47.2	52.0
	$\mu \neq 0, \beta \neq 0$	97.2	94.7	98.0	98.8	45.7	37.6	47.3	52.3
KPSS	$\mu \neq 0, \beta = 0$	23.4	22.0	22.4	25.9	83.4	76.4	82.5	91.3
	$\mu \neq 0, \beta \neq 0$	11.9	7.5	12.4	15.8	70.0	50.5	70.1	89.3

Table 6.2: Rejection rates for seasonal unit root tests (%, $\alpha = 0.01$, Δ = differenced series)

		$D=0$				$D=1$			
Test	Model (6.8)	All	5-year	10-year	20-year	All	5-year	10-year	20-year
ADHF$_\tau$	$\mu = 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}$	85.2	82.2	86.0	86.7	58.0	54.3	59.7	59.9
	$\mu = 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}, \Delta$	96.0	94.5	96.8	96.8	72.0	62.9	74.4	78.6
	$\mu = 0, \beta = 0, \boldsymbol{\gamma} \neq \mathbf{0}$	64.2	50.4	69.6	72.7	45.7	36.6	49.6	51.1
	$\mu = 0, \beta = 0, \boldsymbol{\gamma} \neq \mathbf{0}, \Delta$	82.6	60.8	91.9	95.1	68.2	51.4	76.9	76.2
GM	$\mu \neq 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}$	95.6	96.4	96.2	95.1	9.9	23.4	5.2	0.9
	$\mu \neq 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}, \Delta$	96.2	96.8	96.2	95.7	19.3	36.7	17.6	3.5
HEGY	$\mu \neq 0, \beta \neq 0, \boldsymbol{\gamma} \neq \mathbf{0}$	93.8	81.6	99.8	99.9	89.1	72.8	96.6	98.0
	$\mu \neq 0, \beta \neq 0, \boldsymbol{\gamma} \neq \mathbf{0}, \Delta$	91.8	76.1	99.5	99.9	87.2	74.2	92.9	94.3
	$\mu \neq 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}$	99.4	98.2	99.9	100.0	83.9	69.7	89.0	93.1
	$\mu \neq 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}, \Delta$	98.5	95.8	99.9	100.0	87.9	77.0	91.6	95.0
	$\mu \neq 0, \beta = 0, \boldsymbol{\gamma} \neq \mathbf{0}$	95.0	85.2	99.8	100.0	91.6	88.7	97.6	98.3
	$\mu \neq 0, \beta = 0, \boldsymbol{\gamma} \neq \mathbf{0}, \Delta$	92.8	78.8	99.6	99.9	87.8	76.3	92.9	94.3
	$\mu \neq 0, \beta \neq 0, \boldsymbol{\gamma} = \mathbf{0}$	99.0	97.1	99.9	99.9	82.4	64.3	87.9	94.9
	$\mu \neq 0, \beta \neq 0, \boldsymbol{\gamma} = \mathbf{0}, \Delta$	99.0	97.1	99.9	99.9	82.4	64.3	87.9	94.9
OCSB	$\mu = 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}$	92.2	85.3	94.5	96.9	60.1	46.1	62.5	71.6
	$\mu = 0, \beta = 0, \boldsymbol{\gamma} = \mathbf{0}, \Delta$	84.6	71.9	87.5	94.3	46.6	32.5	47.8	59.5

Table 6.3: Misclassification rates (MCR) of unit root tests (%, $\alpha = 0.01$)

Non-seasonal unit root tests			Seasonal unit root tests		
Test	Model (6.3)	MCR	Test	Model (6.3)	MCR
ADF_τ	$\mu = 0,\ \beta = 0$	46.7	ADHF_τ	$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0}$	36.4
	$\mu \neq 0,\ \beta = 0$	23.4		$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0},\ \Delta$	38.0
	$\mu \neq 0,\ \beta \neq 0$	21.9		$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} \neq \mathbf{0}$	40.8
ERS_τ	$\mu \neq 0,\ \beta = 0$	35.8		$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} \neq \mathbf{0},\ \Delta$	42.8
	$\mu \neq 0,\ \beta \neq 0$	33.8	GM	$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0}$	7.1
ERS_ρ	$\mu \neq 0,\ \beta = 0$	28.4		$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0},\ \Delta$	11.5
	$\mu \neq 0,\ \beta \neq 0$	31.4	HEGY	$\mu \neq 0,\ \beta \neq 0,\ \boldsymbol{\gamma} \neq \mathbf{0}$	47.7
GM	$\mu \neq 0,\ \beta = 0$	15.8		$\mu \neq 0,\ \beta \neq 0,\ \boldsymbol{\gamma} \neq \mathbf{0},\ \Delta$	47.7
PP_ρ	$\mu \neq 0,\ \beta = 0$	15.2		$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0}$	42.3
	$\mu \neq 0,\ \beta \neq 0$	15.2		$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0},\ \Delta$	44.7
PP_τ	$\mu \neq 0,\ \beta = 0$	21.3		$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} \neq \mathbf{0}$	48.3
	$\mu \neq 0,\ \beta \neq 0$	20.7		$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} \neq \mathbf{0},\ \Delta$	47.5
ZA	$\mu \neq 0,\ \beta = 0$	22.8		$\mu \neq 0,\ \beta \neq 0,\ \boldsymbol{\gamma} = \mathbf{0}$	41.7
	$\mu = 0,\ \beta \neq 0$	24.1		$\mu \neq 0,\ \beta \neq 0,\ \boldsymbol{\gamma} = \mathbf{0},\ \Delta$	41.7
	$\mu \neq 0,\ \beta \neq 0$	24.3	OCSB	$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0}$	33.9
KPSS	$\mu \neq 0,\ \beta = 0$	20.0		$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0},\ \Delta$	31.0
	$\mu \neq 0,\ \beta \neq 0$	21.0			

Table 6.3 shows the overall misclassification rates of the non-seasonal and seasonal unit root tests. Whilst especially the PP_ρ and the GM tests perform well in comparison to some of the other tests, they still misclassify about 15% of the time series. This clearly leaves room for improvement by combination of tests, even more so as no individual test dominates the others. For detecting the absence or presence of seasonal unit roots, the GM test—in both variants—clearly outperforms the other tests considered here.

We now seek to find a combination of unit root tests that has lower misclassification rates than any individual test. To this end, we run the CRFE algorithm with 50 conditional RFs in each round. The use of conditional RFs is suggested by the empirical cross-correlations between the p-values of the individual tests, as those range from -0.80 to 0.91 for the non-seasonal unit root tests and from -0.40 to 0.97 for the seasonal unit root tests. To reduce the computational burden, the initial set of tests only includes tests with misclassification rates less than 50% for $d = 0$ and $d = 1$ in the non-seasonal case and for $D = 0$ and $D = 1$ in the seasonal case, resulting in a set of 12 non-seasonal and 4 seasonal unit root tests. Also, the conditional RF in the i-th round is grown on N_i time series, where $N_i = \max\{1000 - (k_i - 1) \times 100, 200\}$ with k_i being the number of tests considered in the i-th round.

In general, due to the random selection of predictors, unit root tests that do not contribute to, or even worsen, the performance are eliminated during the early rounds. The overall misclassification rate is reduced by leaving out these tests and, in most

Table 6.4: Recursive feature elimination for non-seasonal unit root tests

	Elimination after round i		Misclassification rates		
Round i	Test	Model (6.3)	Mean	SD	N_i
1	ADF_τ	$\mu \neq 0,\ \beta \neq 0$	11.1	0.8	200
2	PP_τ	$\mu \neq 0,\ \beta = 0$	11.3	0.8	200
3	ZA	$\mu \neq 0,\ \beta \neq 0$	11.1	0.7	200
4	ZA	$\mu = 0,\ \beta \neq 0$	11.3	0.9	200
5	ADF_τ	$\mu \neq 0,\ \beta = 0$	10.7	0.7	300
6	ZA	$\mu \neq 0,\ \beta = 0$	10.4	0.5	400
7	PP_ρ	$\mu \neq 0,\ \beta = 0$	10.4	0.6	500
8	PP_τ	$\mu \neq 0,\ \beta \neq 0$	10.1	0.6	600
9	KPSS	$\mu \neq 0,\ \beta = 0$	10.1	0.5	700
10	KPSS	$\mu \neq 0,\ \beta \neq 0$	10.4	0.5	800
11	GM	$\mu \neq 0,\ \beta = 0$	12.3	1.2	900
12	PP_ρ	$\mu \neq 0,\ \beta \neq 0$	12.3	0.5	1000

Table 6.5: Recursive feature elimination for seasonal unit root tests

	Elimination after round i		Misclassification rates		
Round i	Test	Model (6.8)	Mean	SD	N_i
1	ADHF_τ	$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} \neq \mathbf{0}$	6.4	0.3	700
2	OCSB	$\mu = 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0},\ \Delta$	6.5	0.4	800
3	GM	$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0},\ \Delta$	6.6	0.4	900
4	GM	$\mu \neq 0,\ \beta = 0,\ \boldsymbol{\gamma} = \mathbf{0}$	7.3	0.2	1000

rounds, the standard deviation of these rates, calculated over the 50 conditional RFs, is decreased. Eventually, the most important tests remain in the set and eliminating more tests increases the overall misclassification rate. As this lower bound is reached, the optimal set of tests is identified.

Table 6.4 reports the CRFE results for non-seasonal unit root tests, indicating that a minimum misclassification rate of almost 10% is achieved when four tests remain in the set: the PP_ρ test with a trend, the GM algorithm and the KPSS tests with a constant and a trend. Thus, the classification performance is increased by more than 5 percentage points compared to the best individual non-seasonal unit root test (PP_ρ tests and GM algorithm, see Table 6.3). Table 6.5 reports the CRFE results for the seasonal unit root tests. The initial set of tests already achieves a misclassification rate of almost 6%. The CRFE approach does not improve this initial benchmark, still we are able to improve about 1 percentage point over the best individual test by combining only four tests.

6.6 Summary

We set out to evaluate the size and power of a large set of unit root tests and their capabilities of correctly identifying non-seasonal and seasonal unit roots. To this end, we simulate time series that are representative of the economic time series in the M4 competition. Furthermore, we employ a conditional random forest-based elimination algorithm to assess which combination of tests decreases the misclassification rates the most.

The best individual unit root tests misclassifies the absence or presence of a non-seasonal unit root in more than 15% of all cases. By combining the p-values of not more than eight unit root tests, the misclassification rate can be reduced to almost 10%. In the case of seasonal unit roots, the reduction is less pronounced. Still, a combination of tests slightly improves the performance of the best individual test.

For future research, we aim to devise an overall unit root test, possibly in combination with a random forest-based overall seasonality test as developed by Webel and Ollech (2018). In this regard, we could also consider a more nuanced lag length selection for the augmented tests studied here as well as additional unit root tests in order to improve the design of our approach. This strategy could then be compared with recent bootstrap approaches to unit root testing (Hansen and Racine, 2018; Palm et al., 2008; Park, 2003; Psaradakis, 2001; Swensen, 2003). Future research could also aim to gain further insights into the theoretical properties of our approach. Some results on the consistency of random forests are already available (Biau et al., 2008), but, to the best of our knowledge, those have not been demonstrated for conditional random forests and the CRFE algorithm so far.

Bibliography

Ayat, L. and Burridge, P. (2000). Unit root tests in the presence of uncertainty about the non-stochastic trend. *Journal of Econometrics*, 95(1):71–96.

Beaulieu, J. J. and Miron, J. A. (1993). Seasonal unit roots in aggregate us data. *Journal of econometrics*, 55(1-2):305–328.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of Random Forest and Other Averaging Classifiers. *Journal of Machine Learning Research*, 9:2015–2033.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Cario, M. C. and Nelson, B. L. (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Clements, M. P. and Hendry, D. F. (2001). Forecasting with difference-stationary and trend-stationary models. *The Econometrics Journal*, 4(1):1–19.

Cochrane, J. H. (1991). A critique of the application of unit root tests. *Journal of Economic Dynamics and Control*, 15(2):275–284.

DeJong, D. N., Nankervis, J. C., Savin, N. E., and Whiteman, C. H. (1992a). Integration versus trend stationary in time series. *Econometrica: Journal of the Econometric Society*, pages 423–433.

DeJong, D. N., Nankervis, J. C., Savin, N. E., and Whiteman, C. H. (1992b). The power problems of unit root test in time series with autoregressive errors. *Journal of Econometrics*, 53(1-3):323–343.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: journal of the Econometric Society*, pages 1057–1072.

Dickey, D. A., Hasza, D. P., and Fuller, W. A. (1984). Testing for unit roots in seasonal time series. *Journal of the American statistical Association*, 79(386):355–367.

Diebold, F. X. and Rudebusch, G. D. (1991). On the power of dickey-fuller tests against fractional alternatives. *Economics letters*, 35:155–160.

Elliott, G. and Müller, U. K. (2006). Minimizing the impact of the initial condition on testing for unit roots. *Journal of Econometrics*, 135(1-2):285–310.

Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica*, 64(4):813–836.

Ernst, P. A., Shepp, L. A., and Wyner, A. J. (2017). Yule‘s "nonsense correlation" solved! *The Annals of Statistics*, 45(4):1789 – 1809.

Franses, P. H. (1991). Seasonality, non-stationarity and the forecasting of monthly time series. *International Journal of forecasting*, 7(2):199–208.

Fuller, W. A. (2009). *Introduction to statistical time series*. John Wiley & Sons.

Gómez, V. and Maravall, A. (2001). Automatic Modeling Methods for Univariate Series. In Peña, D., Tiao, G. C., and Tsay, R. S., editors, *A Course in Time Series Analysis*, pages 171–201. Wiley, New York.

Granger, C. W. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of econometrics*, 2(2):111–120.

Hansen, B. and Racine, J. (2018). Bootstrap model averaging unit root inference. *McMaster University-Department of Economics Working Paper*, 2018(09).

Harvey, D. I., Leybourne, S. J., and Taylor, A. R. (2009). Simple, robust, and powerful tests of the breaking trend hypothesis. *Econometric Theory*, 25(4):995–1029.

Hassler, U. and Hosseinkouchack, M. (2022). Understanding nonsense correlation between (independent) random walks in finite samples. *Statistical Papers*, 63(1):181–195.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hylleberg, S., Engle, R. F., Granger, C. W., and Yoo, B. S. (1990). Seasonal integration and cointegration. *Journal of Econometrics*, 44(1-2):215–238.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1):159 – 178.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74.

Müller, U. K. and Elliott, G. (2003). Tests for unit roots and the initial condition. *Econometrica*, 71(4):1269–1286.

Ng, S. and Perron, P. (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, 69(6):1519–1554.

Ollech, D. and Webel, K. (2020). A random forest-based approach to identifying the most informative seasonality tests. Discussion Paper No 55/2020, Deutsche Bundesbank.

Osborn, D. R., Chui, A. P., Smith, J. P., and Birchenhall, C. R. (1988). Seasonality and the Order of Integration for Consumption. *Oxford Bulletin of Economics and Statistics*, 50(4):361–377.

Palm, F. C., Smeekes, S., and Urbain, J.-P. (2008). Bootstrap unit-root tests: comparison and extensions. *Journal of Time Series Analysis*, 29(2):371–401.

Park, J. Y. (2003). Bootstrap unit root tests. *Econometrica*, 71(6):1845–1895.

Perron, P. (1988). Trends and random walks in macroeconomic time series: Further evidence from a new approach. *Journal of economic dynamics and control*, 12(2-3):297–332.

Phillips, P. C. (1986). Understanding spurious regressions in econometrics. *Journal of econometrics*, 33(3):311–340.

Phillips, P. C. (1987). Time series regression with a unit root. *Econometrica: Journal of the Econometric Society*, pages 277–301.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Psaradakis, Z. (2001). Bootstrap tests for an autoregressive unit root in the presence of weakly dependent errors. *Journal of Time Series Analysis*, 22(5):577–594.

Rodrigues, P. M. M. and Osborn, D. R. (1999). Performance of Seasonal Unit Root Tests for Monthly Data. *Journal of Applied Statistics*, 26(8):985–1004.

Said, S. E. and Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, 71(3):599–607.

Schwert, G. W. (1989). Tests for unit roots: A monte carlo investigation. *Journal of Business & Economic Statistics*, 7(2):147–159.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial Splines and their Tensor Products in Extended Linear Modeling. *The Annals of Statistics*, 25(4):1371–1470.

Swensen, A. R. (2003). Bootstrapping unit root tests for integrated processes. *Journal of Time Series Analysis*, 24(1):99–126.

Webel, K. and Ollech, D. (2018). An overall seasonality test based on recursive feature elimination in conditional random forests. In *Proceedings of the 5th International Conference on Time Series and Forecasting*, pages 20–31. Granada.

West, K. D. (1987). A note on the power of least squares tests for a unit root. *Economics Letters*, 24(3):249–252.

Wolters, J. and Hassler, U. (2006). Unit root testing. *Allgemeines Statistisches Archiv*, 90:43–58.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *Journal of the royal statistical society*, 89(1):1–63.

Zivot, E. and Andrews, D. W. K. (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of business & economic statistics*, 20(1):25–44.

CHAPTER 7

The flexible QS test: A seasonality test for weekly and other non-integer type seasonality

Daniel Ollech

Abstract

Although various seasonality tests exist, none of these tests can handle time series with a non-integer periodicity, such as weekly time series with 52.18 observations per year on average. Here, we generalise the QS test– a variant of the Ljung-Box-test for seasonal autocorrelations developed by Maravall (2012) – to the case of non-integer seasonality. A simulation study shows that this flexibilised variant improves the identification of non-seasonal and seasonal time series, especially in the case of smaller and medium-sized periodicities. The test is also applied to the German truck toll mileage index.

Keywords: Non-integer seasonality; high frequency; parametric tests; non-isochronous; attenuation

7.1 Motivation

The increase in demand for higher frequency economic data has been accelerated by the COVID-19 pandemic (see Ollech, 2022, for an overview of newly established and devised higher frequency indicators). Many of these time series potentially exhibit periodic and seasonal behaviour. To investigate the seasonal nature of a given series, statistical tests are used to identify seasonal and periodic patterns. Furthermore, such tests can be used as a quality measure, e.g. to evaluate whether a seasonally adjusted series does contain residual seasonality (Findley et al., 2017; Ladiray et al., 2018; Lytras et al., 2007).

Although various seasonality tests exist, to the best of our knowledge, none of these tests are flexible enough to handle time series with a non-integer cycle length, such as weekly time series that contain 52.18 observations per year on average.

The QS test relies on estimated seasonal autocorrelations to identify whether a time series is seasonal or not. It can therefore handle both stable and slowly changing seasonal patterns. For monthly time series, it has been shown that the QS test reliably identifies seasonal time series – it is indeed the best performing single seasonality test for this task among all of the tests implemented in the seasonal adjustment software JDemetra+ – and it is subject to only small-sized distortions, if the series has been stationarised appropriately (Ollech and Webel, 2022). Lytras (2015) shows that the QS test performs at least as well as other well-known seasonality tests for quarterly time series. We can therefore plausibly assume it will perform well for other frequencies as well.

In the next section, we will show how the QS test can be generalised to handle non-integer seasonality. section 7.3 analyses the asymptotic behaviour of the test statistic. Section 7.4 uses simulated time series to investigate the performance of the generalised test and section 7.5 applies the test to a set of real world economic time series, while section 7.6 concludes.

7.2 Adapting the QS test

7.2.1 Traditional QS test

The traditional QS test was developed by Maravall (2012). For a seasonal cycle of length τ, it checks whether a weakly stationary time series $\{Y_t\}$ shows significant positive seasonal autocorrelations[1], based on the test statistic

$$QS_{trad} = T\,(T+2)\left(\frac{\hat{\rho}(Y_t, Y_{t-\tau})^2}{T-\tau} + \frac{\left[\max\left\{0, \hat{\rho}(Y_t, Y_{t-2\tau})\right\}\right]^2}{T-2\tau}\right)\mathbb{1}_{\{\hat{\rho}(Y_t, Y_{t-\tau})>0\}}, \tag{7.1}$$

[1] Negative autocorrelation at the seasonal lag implies a year-to-year alternating pattern, which is inconsistent with the definition of seasonality given in Eurostat (2024).

where $\hat{\rho}(l)$ is the lag-l autocorrelation and $\mathbb{1}_{\{.\}}$ is the indicator function of the event in braces and T is the number of observations. Thus, the traditional QS test restricts the Q test for serial dependence in time series developed by Ljung and Box (1978) to the seasonal case.

Maravall (2012) states that the test statistic approximately follows a $\chi^2(2)$-distribution in the case of quarterly and monthly series.[2] To be precise, this is only true for cases where $\hat{\rho}(Y_t, Y_{t-\tau}) > 0$ and, as we will see, does not hold for small sample sizes.

7.2.2 Adaptation of the test statistic

The QS test is based on seasonal autocorrelations. However, the τ-th autocorrelation cannot be computed if τ is non-integer. Instead, we can estimate the correlation between Y_t and the unobservable $Y_{t-\tau}$ through the manifest quasi-observation $Y^{\star}_{t-\tau}$. The latter is defined as

$$Y^{\star}_{t-\tau} := \beta Y_{t-\lfloor\tau\rfloor} + (1-\beta)Y_{t-\lceil\tau\rceil}. \tag{7.2}$$

with floor and ceiling operators defined so that $\lfloor x\rfloor$ is the greatest integer smaller or equal to x and $\lceil x\rceil$ is the least integer greater than or equal to x. The weight β indicates the seasonal portion included in each observation and is thus given by

$$\beta = \tau - \lfloor\tau\rfloor. \tag{7.3}$$

Defined in this way, $Y^{\star}_{t-\tau}$ includes all observations at least partly containing $Y_{t-\tau}$, yet it also includes more than the seasonal signal, i.e.

$$Y^{\star}_{t-\tau} = \beta(\beta Y_{t-\tau} + (1-\beta)\varepsilon_{t-\lfloor\tau\rfloor}) + (1-\beta)([1-\beta]Y_{t-\tau} + \beta\varepsilon_{t-\lceil\tau\rceil}) \tag{7.4}$$

where ε is the part of the observation that is not the seasonal signal, i.e. the non-overlapping share of the current observation and the last periods observation. As an example, assume a weekly series: leaving aside leap years, the x^{th} week in year 2 will contain 6 of the 7 days that fell into the x^{th} week in year two, thus, there is an overlap of 6 days.

Under the null hypothesis of the QS test, $\mathbb{E}(\varepsilon_t) = 0$ and $\mathbb{E}(Y_t\,\varepsilon_t) = 0\,\forall\,t$. Further, the variance of Y_t is given by $\mathbb{V}(Y_t) = \sigma^2_y\;\forall\,t$.

Naively, we might be content with including $Y^{\star}_{t-\tau}$ and $Y^{\star}_{t-2\tau}$ in Equation 7.1. But because $Y^{\star}_{t-\tau}$ includes not only the variable of interest $Y_{t-\tau}$ but also ε, the relationship between Y_t and $Y_{t-\tau}$ will be diluted, similar to the case of a measurement error in cross-sectional data (see Spearman, 1904; Saccenti et al., 2020; Meijer et al., 2021).

Instead, the test statistic needs to be adapted as follows:

[2]A discussion of the asymptotic distribution of the traditional QS test in the case where $\rho(Y_t, Y_{t-\tau}) = 0 \,\&\, \rho(Y_t, Y_{t-2\tau}) = 0$ can be found in Appendix B of McElroy and Roy (2022).

Theorem 1. [Flexible QS]

$$QS_{flex} = T\,(T+2)\left(\frac{(\frac{\hat{\rho}(Y_t,Y^{\star}_{t-\tau})}{A(\beta)})^2}{T-\lceil\tau\rceil} + \frac{\left[\max\{0,\frac{\hat{\rho}(Y_t,Y^{\star}_{t-2\tau})}{A(\beta)}\}\right]^2}{T-\lceil 2\tau\rceil}\right)\mathbb{1}_{\{\hat{\rho}(Y_t,Y^{\star}_{t-\tau})>0\}}, \tag{7.5}$$

where $A(\beta) := \sqrt{(\beta^2+(1-\beta)^2)}$.

Note that $A(\beta)=1$ if $\beta=0$ or $\beta=1$ (see Figure 7.1). Thus, the flexible variant of the QS test is identical to the traditional QS if τ is an integer number.

Proof. Without loss of generality, we assume zero means for the time series, i.e.

$$\mathbb{E}(Y_t) = \mathbb{E}(Y^{\star}_{t-\tau}) = \mathbb{E}(Y_{t-\tau}) = 0.$$

We further assume a stationarised time series[3], e.g. using the first differences of the time series, and – after this differencing – zero first-order autocovariance, i.e. $Cov(Y_t, Y_{t-1}) = 0 \;\forall\, t$.

Then, the autocovariance between Y_t and $Y^{\star}_{t-\tau}$, i.e. $Cov(Y_t, Y^{\star}_{t-\tau})$, in terms of $\mathbb{E}(Y_tY_{t-\tau})$ is given by

$$\begin{aligned}
Cov(Y_t, Y^{\star}_{t-\tau}) &= \mathbb{E}(Y_tY^{\star}_{t-\tau}) - \mathbb{E}(Y_t)\mathbb{E}(Y^{\star}_{t-\tau})\\
&= \mathbb{E}(Y_t[\beta Y_{t-\lfloor\tau\rfloor} + (1-\beta)Y_{t-\lceil\tau\rceil}])\\
&= \beta\mathbb{E}(Y_tY_{t-\lfloor\tau\rfloor}) + (1-\beta)\mathbb{E}(Y_tY_{t-\lceil\tau\rceil})\\
&= \beta\mathbb{E}(Y_t[\beta Y_{t-\tau} + (1-\beta)\varepsilon_{t-\lfloor\tau\rfloor}]) + (1-\beta)\mathbb{E}(Y_t[(1-\beta)Y_{t-\tau} + \beta\varepsilon_{t-\lceil\tau\rceil}])\\
&= \beta^2\mathbb{E}(Y_tY_{t-\tau}) + \beta(1-\beta)\mathbb{E}(Y_t\varepsilon_{t-\lfloor\tau\rfloor}) + (1-\beta)^2\mathbb{E}(Y_tY_{t-\tau}) + \beta(1-\beta)\mathbb{E}(Y_t\varepsilon_{t-\lceil\tau\rceil})\\
&= \beta^2\mathbb{E}(Y_tY_{t-\tau}) + (1-\beta)^2\mathbb{E}(Y_tY_{t-\tau})\\
&= (\beta^2 + (1-\beta)^2)\mathbb{E}(Y_tY_{t-\tau}).
\end{aligned}$$

The variance of the observable $Y^{\star}_{t-\tau}$ is given by

$$\begin{aligned}
\mathbb{V}(Y^{\star}_{t-\tau}) &= \mathbb{V}(\beta Y_{t-\lfloor\tau\rfloor} + (1-\beta)Y_{t-\lceil\tau\rceil})\\
&= \mathbb{V}(\beta Y_{t-\lfloor\tau\rfloor}) + \mathbb{V}([1-\beta]Y_{t-\lceil\tau\rceil}) + 2Cov(\beta Y_{t-\lfloor\tau\rfloor}, (1-\beta)Y_{t-\lceil\tau\rceil})\\
&= \beta^2\sigma_y^2 + (1-\beta)^2\sigma_y^2 + 2Cov(\beta Y_{t-\lfloor\tau\rfloor}, (1-\beta)Y_{t-\lceil\tau\rceil})\\
&= (\beta^2 + (1-\beta)^2)\sigma_y^2.
\end{aligned}$$

[3] The stationarised values are obtained by appropriately differencing the data (see Lytras, 2015; Maravall, 2012). As Ollech and Webel (2022) discuss, the traditional QS test is very sensitive to underdifferencing and untreated serial dependence. The assumption that $Cov(Y_t, Y_{t-1}) = 0 \;\forall\, t$ merely explicitly formulates this sensitivity that is already a feature of the traditional QS test (see also the discussion of the QS test in McElroy and Roy, 2022).

The correlation between Y_t and $Y^\star_{t-\tau}$ is calculated as

$$\begin{aligned}\rho(Y_t, Y^\star_{t-\tau}) &= \frac{Cov(Y_t, Y^\star_{t-\tau})}{\sqrt{\mathbb{V}(Y_t)}\sqrt{\mathbb{V}(Y^\star_{t-\tau})}}\\ &= \frac{(\beta^2 + (1-\beta)^2)\mathbb{E}(Y_t Y_{t-\tau})}{\sqrt{\sigma_y^2}\sqrt{(\beta^2+(1-\beta)^2)\sigma_y^2}}\\ &= \frac{(\beta^2 + (1-\beta)^2)\mathbb{E}(Y_t Y_{t-\tau})}{\sqrt{\sigma_y^2}\sqrt{\sigma_y^2}\sqrt{(\beta^2+(1-\beta)^2)}}\\ &= \frac{\sqrt{(\beta^2 + (1-\beta)^2)}\mathbb{E}(Y_t Y_{t-\tau})}{\sqrt{\sigma_y^2}\sqrt{\sigma_y^2}}. \qquad (7.6)\end{aligned}$$

Meanwhile, the correlation between Y_t and $Y_{t-\tau}$ is given by

$$\begin{aligned}\rho(Y_t, Y_{t-\tau}) &= \frac{Cov(Y_t, Y_{t-\tau})}{\sqrt{\mathbb{V}(Y_t)}\sqrt{\mathbb{V}(Y_{t-\tau})}}\\ &= \frac{\mathbb{E}(Y_t Y_{t-\tau}) - \mathbb{E}(Y_t)\mathbb{E}(Y_{t-\tau})}{\sqrt{\sigma_y^2}\sqrt{\sigma_y^2}}\\ &= \frac{\mathbb{E}(Y_t Y_{t-\tau})}{\sqrt{\sigma_y^2}\sqrt{\sigma_y^2}}. \qquad (7.7)\end{aligned}$$

By comparing Equation 7.6 and Equation 7.7 we can derive the attenuation factor

$$\begin{aligned}\rho(Y_t, Y^\star_{t-\tau}) &= \sqrt{(\beta^2+(1-\beta)^2)}\rho(Y_t, Y_{t-\tau})\\ \frac{\rho(Y_t, Y^\star_{t-\tau})}{\sqrt{(\beta^2+(1-\beta)^2)}} &= \rho(Y_t, Y_{t-\tau}),\end{aligned}$$

so that the correction factor needed to obtain the correct autocorrelation is given by

$$A(\beta) := \sqrt{(\beta^2+(1-\beta)^2)}.$$

$\square$

7.3 Asymptotics

As mentioned above, Maravall (2012) states that the QS test approximately follows a $\chi^2_{df=2}$-distribution and shows this based on simulated monthly series with 10 and 20 years of observations.

For any seasonal autocorrelation, Dufour and Roy (1986, p.6) show that

$$\mathbb{E}(\hat{\rho}(Y_t, Y^\star_{t-\tau})^2) = \frac{T^3 - T^2(\tau+1) + 3\tau^2}{T^2(T^2-1)} \qquad (7.8)$$

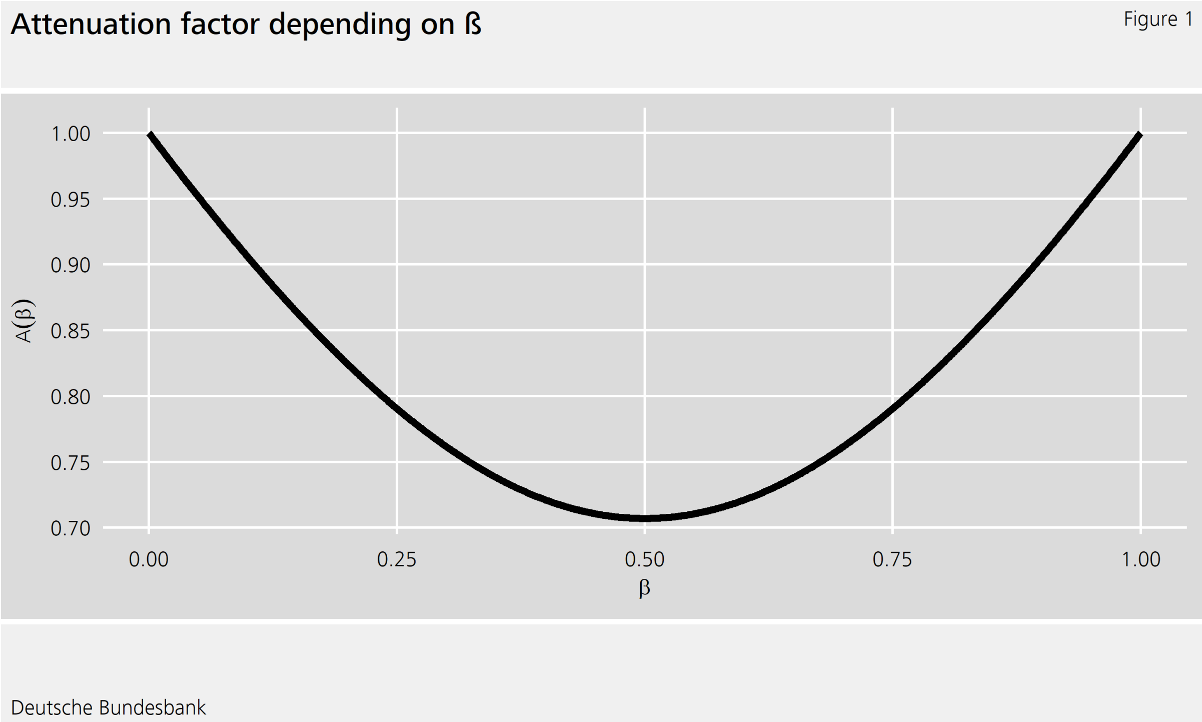

Figure 7.1: Attenuation factor depending on β

We can use this to derive the expected value of the test statistic not restricted to positive seasonality under the null.

$$\begin{aligned}
\mathbb{E}(QS_{flex}) &= \mathbb{E}\left[(T\,(T+2) \left(\frac{(\frac{\hat{\rho}(Y_t, Y^\star_{t-\tau})}{A(\beta)})^2}{T - \lceil \tau \rceil} + \frac{\left[\frac{\hat{\rho}(Y_t, Y^\star_{t-2\tau})}{A(\beta)}\right]^2}{T - \lceil 2\tau \rceil} \right) \right] \\
&= \frac{T(T+2)}{(T-\lceil\tau\rceil)A(\beta)^2}\mathbb{E}\left[\hat{\rho}(Y_t, Y^\star_{t-\tau})^2)\right] + \frac{T(T+2)}{(T-\lceil 2\tau\rceil)A(\beta)^2}\mathbb{E}\left[\hat{\rho}(Y_t, Y^\star_{t-2\tau})^2)\right] \\
&= \frac{T(T+2)}{(T-\lceil\tau\rceil)A(\beta)^2}\frac{T^3 - T^2(\tau+1) + 3\tau^2}{T^2(T^2-1)} \\
&+ \frac{T(T+2)}{(T-\lceil 2\tau\rceil)A(\beta)^2}\frac{T^3 - T^2(2\tau+1) + 12\tau^2}{T^2(T^2-1)} \\
&= \frac{T^4(T+2)}{[(T-\lceil\tau\rceil)A(\beta)^2][T^4-T^2]} - \frac{T^3(T+2)(\tau+1)}{[(T-\lceil\tau\rceil)A(\beta)^2][T^4-T^2]} \\
&+ \frac{3T(T+2)\tau^2}{[(T-\lceil\tau\rceil)A(\beta)^2][T^4-T^2]} + \frac{T^4(T+2)}{[(T-\lceil 2\tau\rceil)A(\beta)^2][T^4-T^2]} \\
&- \frac{T^3(T+2)(2\tau+1)}{[(T-\lceil 2\tau\rceil)A(\beta)^2][T^4-T^2]} + \frac{12T(T+2)\tau^2}{[(T-\lceil 2\tau\rceil)A(\beta)^2][T^4-T^2]} \\
&= 1 + O(T^{-1}) + O(T^{-3}) + 1 + O(T^{-1}) + O(T^{-3}) \\
&= 2 + O(T^{-1})
\end{aligned}$$

Thus it can be seen that the mean converges to the mean of the $\chi^2_{df=2}$-distribution as the time series length increases. This is in line with the investigation of the size distortions presented in subsection 7.A.1.

7.4 Simulations

We set out to evaluate the performance of the QS_{flex} test. In addition to weekly series, i.e. series with $\tau = 365/7 = 52.14$, we include variants thereof, namely two-weekly ($\tau = 365/14 = 26.07$) and eight-weekly ($\tau = 365/56 = 6.52$) series.[4] Some time series have a monthly recurring pattern, so we test a monthly pattern in weekly time series ($\tau = 365/7/12 = 4.35$). Additionally, we include two-daily ($\tau = 7/2 = 3.5$) and three-daily series ($\tau = 7/3 = 2.33$), yielding six different non-integer cycle lengths in total.

The time series used in this simulation exercise are obtained by first simulating daily-time series with an annually recurring seasonal pattern of length 365 and then temporally aggregating these series by averaging to weekly, two-weekly and eight-weekly series.

Additionally, we simulate daily time series with a weekly pattern and analogously use them to obtain the two-daily and three-daily series. Weekly time series with a monthly recurring pattern are obtained by simulating daily time series with a monthly recurring pattern[5] and aggregating them to a weekly series.

We simulate both seasonal and non-seasonal time series. The seasonal series are simulated based on the following time series model

$$Y_t = S_t * T_t * I_t, \tag{7.9}$$

where additive seasonal components, S_t, are simulated as a seasonal random walk with autocorrelated innovations

$$\tilde{S}_t = \tilde{S}_{t-\tau} + \xi_t \tag{7.10}$$

$$\xi_t = \beta\xi_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma^2_\epsilon). \tag{7.11}$$

To obtain the multiplicative seasonal component, $\tilde{S}_t$ is transformed as follows

$$S_t = 1 + \tilde{S}_t/100. \tag{7.12}$$

The standard deviation of the seasonal component after the temporal aggregation, σ_S,controls the magnitude of the seasonality. We include $\sigma_S \in \{2, 5, 10\}$ and $\sigma_\epsilon \in \{0, 0.01, 0.1, 1\}$ for

[4] Note that leap years are omitted.

[5] The monthly recurring pattern is obtained by simulating a seasonal pattern of length 31 according to Equation 7.10 and then omitting the superfluous days.

the set of seasonal time series. For the simulation of the non-seasonal time series, the component S_t in Equation 7.9 is omitted.

Let an ARIMA model of order (p,d,q) be given by

$$\phi_p(B)\,\nabla_1^d Y_t = \theta_q(B)\,\varepsilon_t. \tag{7.13}$$

where B is the backshift operator, $\nabla^d = 1 - B^d$, d indicates the order of differencing, ϕ_p is an autoregressive polynomials of order p, θ_q is a non-seasonal moving average polynomial of order q, and $\{\varepsilon_t\}$ are white noise innovations.

The trend-irregular components, $T_t * I_t$, are simulated jointly from a broad range of pre-defined ARIMA(p,d,q) models.[6]

We consider four time series length, $T \in \{5\tau, 10\tau, 20\tau, 50\tau\}$. For each combination of parameters, we simulate 150 seasonal resp. 1800 non-seasonal time series, to get a total of $50,400 \times 6$ non-seasonal and $50,400 \times 6$ seasonal time series, respectively.

We compare the flexible variant of the QS test, that takes into account the non-integer with the traditional variant, where τ is rounded to the next integer.[7] For time series with less than 150 observations, the null distribution is given by Monte Carlo simulation; for longer time series the $\chi^2_{df=2}$-distribution is employed (cf. subsection 7.A.1).

Table 7.1 shows the rejection rates of the flexible and traditional variants of the QS test. For seasonal time series, the detection rate of the flexible variant of the QS is considerably better for all smaller periodicities. Only for the weekly and two-weekly series is the difference negligible or in favour of the traditional QS test. This is mostly due to time series simulated from more complex ARIMA models, where the flexible version tends to perform slightly worse.[8]

For non-seasonal time series, both variants tend to have an error rate below their respective significance level. Yet, the size of the flexible QS test is closer to its actual significance level in almost all instances.

7.5 Real-world example

The German truck toll mileage index measures the daily distance covered by trucks in Germany. Its seasonal characteristics have been described in detail by Ollech (2022). Most importantly, the series contains both weekly and annually recurring pattern, as can be seen in Figure 7.2, where different temporal aggregations of the time series is displayed. The weekly pattern is discernible even in the series aggregated to a two- and three-daily

[6] The following models are considered: ARIMA(0,0,0), ARIMA(0,1,1) with $\theta_q = 0.75$, ARIMA(0,1,1) with $\theta_q = -0.5$, , ARIMA(1,1,0) with $\phi_p = -0.75$, ARIMA(0,1,1) with $\phi_p = 0.5$, ARIMA(3,1,1) with $\theta_q = 0.8$ and $\phi_p = (0.3, -0.1, 0.2)$, ARIMA(0,1,2) with $\theta_q = (0.5, 0.3)$.

[7] The IEC 60559 standard is used for rounding, so that 3.5 is rounded to 4.

[8] Detailed simulation results can be obtained from the author.

Table 7.1: Rejection rates of null hypothesis, in %, for variants of the QS test for different periodicities (τ)

		Seasonal time series		Non-seasonal time series	
	τ	Flexible	Traditional	Flexible	Traditional
$\alpha = 0.01$	2.33	58.30	39.60	2.32	1.48
	3.50	78.16	24.96	1.76	0.82
	4.35	92.60	88.25	1.10	0.72
	6.52	80.65	41.29	1.17	0.80
	26.07	96.92	97.04	0.82	0.66
	52.14	99.59	99.69	0.86	0.58
$\alpha = 0.05$	2.33	72.86	55.35	5.47	4.63
	3.50	78.89	34.79	4.36	2.92
	4.35	95.31	92.90	3.55	2.72
	6.52	84.57	54.99	3.71	2.86
	26.07	97.94	98.05	2.98	2.73
	52.14	99.79	99.86	3.03	2.35

Note: Two variants of the QS test are included. The flexible variant takes into account the non-integer nature of the frequency while the traditional variant is calculated with the periodicity rounded to the closest integer.

series. The annual pattern is clearly visible in the weekly series, but considerably less pronounced in the eight-weekly series.

Table 7.2 applies the QS_{flex} test to the truck toll mileage time series aggregated to different periodicities and also includes the traditional QS test – where the frequency rounded to the next integer is used – as a comparison.

In all time series variants considered, the test statistic is higher for the flexible version, showing stronger evidence for seasonality. In almost all cases, the test results are highly significant. Only for the series aggregated to an eight-weekly series is there a practically relevant difference. The flexible QS test finds significant evidence of a seasonal pattern, while the traditional QS variant does not find significant seasonality at the 5% significance level.

Thus, the results are consistent with the visual inspection: in all instances of a clearly visible seasonal pattern, both variants of the QS test are able to identify the seasonality correctly. When the seasonal pattern is less pronounced, only the flexible variant of the QS test is able to detect the seasonal pattern.

Table 7.2: Test results of variants of the QS test for the German truck toll mileage index aggregated to different periodicities (τ)

	Three-daily	Two-daily	Eight-weekly	Weekly
	$\tau = 2.33$	$\tau = 3.50$	$\tau = 6.52$	$\tau = 52.18$
Flexible QS	614.5***	3404.5***	71.3***	662.9***
Traditional QS	573.4***	573.2***	5.9	574.9***

Note: * p < .05, ** ≤ .01, *** p ≤ .001.

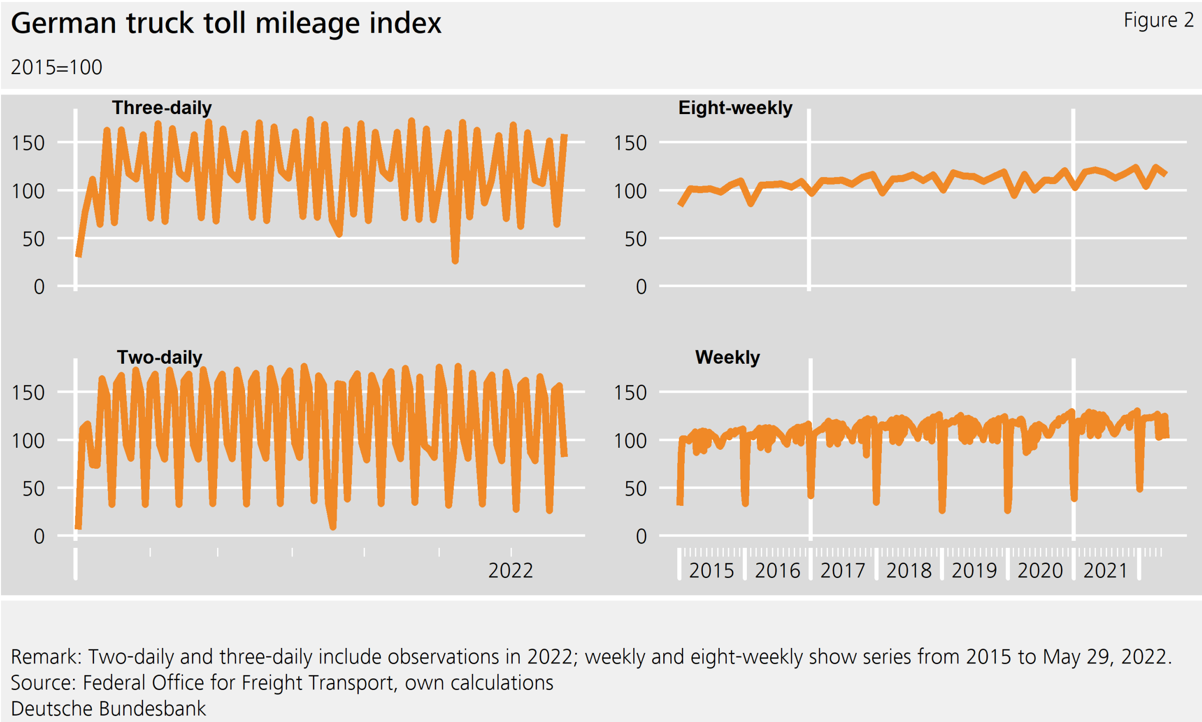

Figure 7.2: German truck toll mileage index

7.6 Summary

We generalised the *QS* test to allow for non-integer lengths of the seasonal pattern. A simulation study showed that – over all periodicities considered – this flexibilisation improves the identification of non-seasonal and seasonal time series, especially for shorter, non-integer cycle lengths.

Ollech and Webel (2022) show how a combination of seasonality tests can improve upon the accuracy of the individual tests. Especially combinations that include tests for different manifestions of seasonality are worth considering. As the *QS* test allows for changing seasonality, it may be attractive to combine it with tests that are designed to detect stable seasonality, such as the *Friedman* test or the *F-test on seasonal dummies* (for details on these tests, see Ollech and Webel, 2022). Thus, future research may try to generalise the *Friedman* or a similar test to allow for non-integer seasonalities.

7.A Appendix

7.A.1 Size distortions

Here we present simulation results that investigate the size distortions of the *QS* test. To this end, we show how the proximity between the empirical and theoretical null distribution converge for different periodicities and time series length. One way to analyse whether there are significant size distortions is to use the Kolmogorov-Smirnov test and the test developed by Anderson and Darling (1954). However, size distortions at conventional significance levels are usually of greater concern than other local deviations from the theoretical distribution. Therefore, the main focus lies on size distortions at the 1% and 5% significance level.

Table 7.3 includes results for all of the non-integer cycle lengths mentioned above as well as common integer periodicities, namely bi-annual ($\tau = 2$), quarterly ($\tau = 4$), and monthly ($\tau = 12$) and the day-of-the-week effect ($\tau = 7$) in daily time series.

The results are obtained by simulating white noise series of length $n = \tau * \textit{number of cycles}$ rounded to the next integer and computing the test statistic according to Equation 7.5 on the series without differencing.

Generally, the QS test shows size distortions at the 1% and 5% level for all but long time series. For short time series, the test tends to be slightly undersized for integer frequencies, while it is oversized for non-integer frequencies with $\tau <= 6.52$, most noticeably for three-daily series. In these cases, using the $\chi^2_{df=2}$-distribution might not be advisable. Instead, a null distribution obtained by Monte Carlo simulation may be used.

The Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit test compare the empirical p-values to a uniform distribution.[9] These tests detect significant deviations for all but very long time series.[10]

For weekly time series there is some statistical indication that the null distribution is not exactly equal to $\chi^2_{df=2}$ for all time series with less than 50 years of observations. However, graphically and with respect to the size distortions at the conventional significance levels, the difference is negligible. For shorter series there may be some statistically significant but practically irrelevant differences.

For time series with 150 or more observations, the size distortions are smaller or equal to 0.2 at the 1% significance level for the frequencies considered here, and smaller or equal to 0.3 at the 5% significance level. As a rule of thumb, the $\chi^2_{df=2}$-distribution may, therefore, be considered a reasonable approximation if the number of observations exceeds 150 for $\alpha = 1\%$ and $\alpha = 5\%$.

[9] Testing whether the p-values are uniformly distributed is equivalent to testing whether the test statistic follows a $\chi^2_{df=2}$-distribution

[10] In the case of $\tau = 7$, i.e. a weekday cycle in a daily time series, even 250 cycles correspond to only about five years of observations, which is not necessarily considered a long time series.

Table 7.3: Size of QS test, in %, depending on frequency (τ) and number of cycles

Level of significance	Number of cycles 5	10	20	30	40	50	75	100	150	250	500	1000
$\tau = 2$												
$\alpha = 0.01$	1.0	0.9	1.1	1.1	1.2	1.2	1.2	1.2	1.1	1.2	1.1	1.1
$\alpha = 0.05$	5.4	4.7	5.1	5.3	5.3	5.4	5.3	5.3	5.2	5.3	5.1	5.1
$\tau = 2.33$												
$\alpha = 0.01$	0.0	0.4	1.0	1.1	1.1	1.1	1.2	1.1	1.1	1.1	1.1	1.0
$\alpha = 0.05$	0.0	2.3	4.3	4.6	4.8	4.8	5.1	5.1	5.1	5.1	5.1	5.0
$\tau = 3.50$												
$\alpha = 0.01$	0.6	0.8	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.0	1.1
$\alpha = 0.05$	3.7	4.1	4.9	5.0	5.0	5.0	5.0	5.0	5.1	5.0	5.0	5.1
$\tau = 4$												
$\alpha = 0.01$	1.2	1.1	1.2	1.1	1.2	1.2	1.2	1.1	1.1	1.1	1.1	1.0
$\alpha = 0.05$	5.6	5.3	5.4	5.3	5.3	5.4	5.2	5.2	5.1	5.2	5.1	5.1
$\tau = 4.35$												
$\alpha = 0.01$	0.8	0.9	1.1	1.1	1.1	1.2	1.1	1.1	1.1	1.1	1.1	1.1
$\alpha = 0.05$	3.9	4.1	4.8	4.9	5.0	5.0	5.0	5.0	5.0	5.0	5.1	5.1
$\tau = 6.52$												
$\alpha = 0.01$	1.0	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.0	1.0	1.0
$\alpha = 0.05$	4.6	4.7	5.1	5.0	5.0	5.1	5.1	5.0	5.0	5.0	5.1	5.0
$\tau = 7$												
$\alpha = 0.01$	1.2	1.2	1.2	1.2	1.2	1.1	1.1	1.1	1.1	1.1	1.1	1.0
$\alpha = 0.05$	5.5	5.4	5.4	5.3	5.3	5.2	5.1	5.2	5.2	5.2	5.1	5.1
$\tau = 12$												
$\alpha = 0.01$	1.2	1.2	1.2	1.1	1.1	1.1	1.1	1.1	1.1	1.0	1.0	1.0
$\alpha = 0.05$	5.6	5.3	5.2	5.2	5.2	5.1	5.2	5.1	5.2	5.0	5.0	5.1
$\tau = 26.07$												
$\alpha = 0.01$	1.3	1.2	1.1	1.0	1.0	1.1	1.1	1.1	1.0	1.0	1.0	1.0
$\alpha = 0.05$	5.4	5.3	5.2	5.1	5.1	5.1	5.0	5.0	5.1	5.0	5.1	5.0
$\tau = 52.14$												
$\alpha = 0.01$	1.1	1.1	1.1	1.1	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\alpha = 0.05$	5.3	5.1	5.2	5.1	5.1	5.0	5.0	5.0	5.0	5.0	5.0	5.0

Note: QS test is applied to time series simulated from a standard normal distribution of length $n = (\tau * \text{number of cycles})$ rounded to next integer. For each frequency τ, 100,000 time series have been simulated. Grey cell colours indicate that both the KS and the AD goodness of fit tests (p-values against uniform distribution) are significant at the 1% level.

Bibliography

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

Dufour, J.-M. and Roy, R. (1986). Generalized portmanteau statistics and tests of randomness. *Communications in Statistics-Theory and Methods*, 15(10):2953–2972.

Eurostat (2024). ESS Guidelines on Seasonal Adjustment. 2024 edition. Publications Office of the European Union, Luxembourg. ISBN 978-92-68-15943-9.

Findley, D. F., Lytras, D. P., and McElroy, T. S. (2017). Detecting Seasonality in Seasonally Adjusted Monthly Time Series. Research Report 2017-03, U.S. Census Bureau.

Ladiray, D., Palate, J., Mazzi, G. L., and Proietti, T. (2018). Seasonal Adjustment of Daily and Weekly Data. In Mazzi, G. L., Ladiray, D., and Rieser, D. A., editors, *Handbook on Seasonal Adjustment*, chapter 29, pages 757–783. Publications Office of the European Union, Luxembourg.

Ljung, G. M. and Box, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 65(2):297–303.

Lytras, D. (2015). On Seasonality: Comparing X-13ARIMA-SEATS Diagnostics for Quarterly Series. In *Proceedings of the 2015 Joint Statistical Meetings*, Seattle.

Lytras, D. P., Feldpausch, R. M., and Bell, W. R. (2007). Determining Seasonality: A Comparison of Diagnostics from X-12-ARIMA. In *Proceedings of the International Conference on Establishment Surveys III. (CD-ROM)*.

Maravall, A. (2012). *Seasonality Tests and Automatic Model Identification in TRAMO-SEATS*. Bank of Spain. Mimeo.

McElroy, T. and Roy, A. (2022). A review of seasonal adjustment diagnostics. *International Statistical Review*, 90(2):259–284.

Meijer, E., Oczkowski, E., and Wansbeek, T. (2021). How measurement error affects inference in linear regression. *Empirical Economics*, 60(1):131–155.

Ollech, D. (2022). Economic analysis using higher-frequency time series: Challenges for seasonal adjustment. *Empirical Economics*, pages 1–24.

Ollech, D. and Webel, K. (2022). A random forest-based approach to combining and ranking seasonality tests. *Journal of Econometric Methods*.

Saccenti, E., Hendriks, M. H., and Smilde, A. K. (2020). Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. *Scientific Reports*, 10(1):1–19.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101.

CHAPTER 8

Conclusion

The seasonal adjustment of a time series is always a multi-step endeavour. The process can be broadly categorised into three distinct phases:

1. **Preliminary investigation:** Initially, the time series must be thoroughly examined. This involves investigating whether calendar effects are present and determining if seasonal effects are detectable. An optimal model for the time series must then be identified, taking into account these characteristics.

2. **Calendar and seasonal adjustment:** Subsequently, the calendar and seasonal adjustment is conducted. This includes the selection of appropriate regressors and the application of suitable methods to effectively estimate and remove the identified calendar and seasonal influences.

3. **Post-adjustment analysis:** Finally, the results of the adjustment must be evaluated. This analysis seeks to uncover any indications of unadjusted calendar effects or residual seasonality that may suggest the need for further refinement of the adjustment process.

The collection of papers presented herein contributes to each of these steps in several ways:

To identify the presence of seasonal effects within a series, seasonality tests are typically utilised. **Chapter Four** evaluates a set of common seasonality tests, while **Chapter Five** introduces a comprehensive seasonality test with a notably low misclassification rate for both seasonal and non-seasonal series. **Chapter Seven** further enhances one of the most reliable seasonality tests to accommodate weekly and other time series characterised by a non-integer length of the seasonal cycle.

The identification of a suitable RegARIMA model is crucial for effective calendar adjustment, and this frequently encompasses unit root testing. **Chapter Six** assesses a broad array of unit root tests and proposes a test combination whose precision surpasses that of any individual test within the set.

Chapters Two and Three delve into the seasonal adjustment of time series with frequencies other than the traditional monthly and quarterly frequencies. These chapters place particular emphasis on higher frequency time series, which have gained increased relevance as analysts and policymakers grappled with evaluating the economic conditions during the different stages of the COVID-19 pandemic.

In summary, the papers contribute to the advancement of statistical methods for seasonal adjustment, providing both theoretical insights and practical tools to address the challenges posed by various types of time series data. The methods discussed and developed are available in designated R packages: {`dsa`}, {`seastests`} and {`tssim`}.